Methods for learning surgical instrument segmentation from unlabelled datasets using prior knowledge

Luca Sestini

UNIVERSITÉ DE STRASBOURG

DOCTORAL SCHOOL MSII

ICube Laboratory (UMR 7357)

THESIS

presented by:

Luca SESTINI

Defended on: **June 8th, 2023**

For obtaining the degree of **Doctor of Philosophy**
from the University of Strasbourg

Field: Image and Vision

Methods for Learning Surgical Instrument Segmentation from Unlabelled Datasets using Prior Knowledge

Thesis directors:	
Prof. Nicolas Padoy	Professor, Université de Strasbourg
Prof. Giancarlo Ferrigno	Professor, Politecnico di Milano
Chair of the Committee:	
Prof. Caroline Essert	Professor, Université de Strasbourg
Reviewers:	
Prof. Diana Mateus	Professor, École Centrale de Nantes
Prof. Danail Stoyanov	Professor, University College of London
Examiners:	
Dr. Elena De Momi	Associate Professor, Politecnico di Milano
Dr. Benoit Rosa	Research Scientist, CNRS
Dr. Mathias Unberath	Assistant Professor, Johns Hopkins University

Abstract

As we are entering an era of rapid technological evolution fueled by big data and data science, surgery is bound to be radically transformed. Today's operating room, the center of surgical care, is a highly technological environment: the variety of signals produced inside it by digital devices can capture the complexity of the surgical act, and provide data to describe it. The analysis of such data through the emerging field of surgical data science offers tangible opportunities to make surgery safer, more efficient and more accessible. Among these data, surgical videos represent the richest and most comprehensive source of information describing the surgical act in minimally invasive procedures. The automatic localisation and identification of surgical instruments from such videos is an essential component of valuable downstream applications like automatic surgical skill assessment and real-time decision support, aimed at facilitating surgical training and at providing intra-operative assistance. Most of the available solutions tackling this problem use fully-supervised learning approaches to train deep learning models on manually annotated data. Due to the cost of annotations, the training of such models is confined to limited sets of labelled and curated data, potentially impacting their generalization ability to perform on real-world data.

This thesis explores methods for learning instrument localisation and identification from unlabelled datasets. To this aim, we first identify several possible sources of information providing general knowledge about surgical instruments. Such knowledge is significantly more cost-effective to obtain compared to standard manual annotations, and easily repurposable across surgical domains. Then, we make several contributions, showing different ways to formalize such knowledge and to inject it in deep learning model architectures to solve the tool localisation and identification problems. Specifically, we tackle the increasingly complex tasks of binary tool segmentation, instance segmentation and image-based 3D pose estimation. All our approaches are trained on completely unlabelled data, by fabricating effective pseudo-supervision signals from prior knowledge and complementary multi-modal data. This is achieved by means of novel methods for unsupervised learning, self-supervised representation learning, and learning from noisy labels, all of which are designed to effectively leverage such prior and complementary knowledge. We hope that our proposed approaches will facilitate the development of valuable assistive technologies to enhance the quality of surgical care.

Acknowledgements

This manuscript is the end result of several years of studies, as part of a joint doctorate between the Université de Strasbourg, as part of the CAMMA lab, and Politecnico di Milano, as part of the NEARLab. Before presenting the research carried out over this period of time, I will use the next few lines to honor the help and support that was given to me throughout my thesis.

First of all, I would like to thank the members of the jury: Pr. Caroline Essert, Pr. Diana Mateus, Pr. Danail Stoyanov, Dr. Elena De Momi and Dr. Mathias Unberath. It was an honor to have my work evaluated by such influential members of the community. A special thanks go to Diana and Mathias: as part of my mid-thesis committee, you have really contributed to shaping my work in the form it is now.

I would like to thank my main supervisors Pr. Nicolas Padoy and Dr. Benoit Rosa: I started this journey with little to no knowledge about the field. You were always present with precious bits of advice while giving me time to make mistakes, learn and finally succeed. I couldn't have asked for better. Thanks to Dr. Elena De Momi and Pr. Giancarlo Ferrigno, my supervisors from the Politecnico di Milano side. You have been my professors during my master's degree in Politecnico, you have supervised my master's thesis and now my doctorate. It was my pleasure to work with you.

Much of this work would not have been possible without a number of key collaborators. In that regard, I would like to thank the whole ATLAS consortium, particularly Dr. Gianni Borghesan and Dr. Emmanuel Vander Poorten for leading the project. Thanks to Bernard Dallemagne, Florent Nageotte, and Philippe Zanne for their help in generating the STRAS dataset, a crucial piece of my work.

I am thankful of course for every member of team CAMMA - past and present. All of them contributed to creating a work environment I will remember very fondly long after my time here; especially Deepak, Pietro, Chinedu, Tong and Vinkle, who have inspired me and helped me throughout these years. Deepak, Pietro, Tong, the best is yet to come!

Outside of the lab, I am definitely thankful for the company of my hometown friends. Every time I come home to you, even after months, it feels like nothing has changed since when we were kids. A heartfelt thanks to Ginevra who has come into my life in the last year of my PhD, the hardest, and, thanks to you, the most beautiful.

Of course, all of this would not be complete without thanking my family. Thanks to my grandparents Gino, Gigi, Popa. You cannot be here to celebrate, but I know you would

be proud of me. *Nonno* Gigi, you have been part of this all journey, thanks for all you love. I miss you all. Thanks *nonna* Nada, your sweetness and intelligence have always been my pillar of strength. Finally *babbo, mamma*, thanks for your continuous support. You were the first ones to help me in the bad times, and the first ones to celebrate with me in the good times. This would never have been possible without you two. *Vi voglio bene.*

Contents

1 Introduction **21**
- 1.1 Background . . . 22
  - 1.1.1 The Big Data Gold Rush . . . 22
  - 1.1.2 Healthcare in the Big Data Era . . . 23
- 1.2 Surgery in the Big Data Era . . . 24
  - 1.2.1 The *datafied* Operating Room in the Minimally Invasive Surgery Era . . . 25
  - 1.2.2 Emerging Challenges in Minimally Invasive Surgery . . . 27
  - 1.2.3 Opportunities for Surgical Data Science . . . 28
  - 1.2.4 The Value of Endoscopic Video Analysis . . . 29
- 1.3 Surgical Computer Vision . . . 30
  - 1.3.1 *Instrument-Centrality* of Surgical Computer Vision tasks . . . 30
  - 1.3.2 From Building-Blocks to Applications . . . 33
- 1.4 Surgical Instrument Localisation and Identification . . . 34
  - 1.4.1 Problem Statement . . . 35
  - 1.4.2 The Annotation Bottleneck . . . 39
  - 1.4.3 The Potential of Prior and Complementary Knowledge . . . 41
  - 1.4.4 Research Question . . . 43
- 1.5 Thesis Contribution . . . 44
- 1.6 Outline . . . 46

2 Related Work on Surgical Tool Localisation and Identification **47**
- 2.1 Fully-Supervised Solutions . . . 48
- 2.2 Semi-Supervised Solutions . . . 48
- 2.3 Prior and Complementary Knowledge based Solutions . . . 49
  - 2.3.1 Simulation and Semi-synthetic Data Generation . . . 50
  - 2.3.2 Complementary Information based Solutions . . . 51
  - 2.3.3 Weak Prior Knowledge based Solutions . . . 52
  - 2.3.4 Strong Prior Knowledge based Solutions . . . 54
- 2.4 Thesis Positioning . . . 56

3 FUN-SIS: a Fully Unsupervised Approach for Surgical Instrument Segmentation: 59
3.1 Introduction . . . 60
3.1.1 Objective & Contributions . . . 60
3.1.2 Learning from Motion and Noisy Labels . . . 61
3.2 Methodology . . . 63
3.2.1 Step 1: Unsupervised Motion Segmentation . . . 64
3.2.2 Step 2: The Proxy Segmentation Network . . . 66
3.2.3 Step 3: Refining Noisy Labels . . . 67
3.2.4 Training Strategy . . . 69
3.3 Experimental Set-up . . . 70
3.3.1 Datasets . . . 70
3.3.2 Artificially Corrupted Datasets . . . 72
3.3.3 Design Choices & Training Details . . . 73
3.4 Experiments and Results Analysis . . . 73
3.4.1 Optical-Flow Segmentation . . . 74
3.4.2 Single-frame Binary Tool segmentation . . . 76
3.5 Ablation Studies and Additional Experiments . . . 81
3.5.1 Optical-Flow Augmentation and Noise Vector Size . . . 81
3.5.2 Proxy Network Architecture . . . 82
3.5.3 Loss Function Coefficients (α_P, α_S) . . . 82
3.5.4 Local IoU Parameters' Impact . . . 83
3.5.5 Shape-Priors Quality & Quantity . . . 85
3.5.6 Noise Properties (Unpredictability & Polarization) . . . 86
3.5.7 Per-class IoU evaluation . . . 89
3.5.8 Random Unlabelled Data . . . 90
3.5.9 FUN-SIS Applicability on another Domain: Cholec80 . . . 90
3.6 Discussion and Future Work . . . 92
3.7 Conclusion . . . 94

4 PAF-IS: a Pixel-wise Annotation Free framework for Instance Segmentation of Surgical Tools 95
4.1 Introduction . . . 96
4.1.1 Objective & Contributions . . . 97
4.2 Methodology . . . 97
4.2.1 Tool Instantiation . . . 98
4.2.2 Instance-wise Feature Representation Learning . . . 101
4.2.3 Instance-wise Tool Type Classification . . . 102
4.3 Experimental Set-up . . . 105
4.3.1 Datasets . . . 105
4.3.2 Design Choices & Training Details . . . 106
4.4 Experiments and Results Analysis . . . 107
4.4.1 Tool Instantiation . . . 107
4.4.2 Tool Instance Segmentation . . . 108
4.5 Ablation Studies . . . 109
4.5.1 Tool Instantiation Augmentation Strategy . . . 110
4.5.2 Tool Instantiation Inference Parameters . . . 111

4.5.3 Prototype Labels Number . . . 111
4.6 Discussion . . . 113
4.7 Conclusion . . . 114

5 KI-BOT: a Kinematic Bottleneck Approach For Pose Regression of Flexible Surgical Instruments directly from Images 117
5.1 Introduction . . . 118
5.1.1 Objective & Contributions . . . 119
5.2 Methodology . . . 119
5.2.1 Method Overview . . . 120
5.2.2 Backgroundizer module: an Inpainting Problem . . . 121
5.2.3 Camera and Robot Modelling . . . 122
5.2.4 Regressor and Decoder . . . 123
5.3 Experimental Set-up . . . 123
5.3.1 Robotic System . . . 123
5.3.2 Datasets . . . 124
5.3.3 Design Choices & Training Details . . . 126
5.4 Experiments and Results Analysis . . . 126
5.4.1 Backgroundizer and Physical Modules Results . . . 127
5.4.2 Kinematic Regression Results . . . 127
5.5 Discussion . . . 129
5.6 Conclusion . . . 132

6 Conclusion 133
6.1 A Unified Framework for Learning from Unlabelled Datasets . . . 134
6.1.1 Framework Contextualization . . . 137
6.2 Discussion and Future Work . . . 138
6.2.1 Limitations . . . 138
6.2.2 Opportunities . . . 139
6.2.3 Open Questions . . . 140

Appendices 143

A List of Publications 145

B Résumé en français 147
B.1 Introduction . . . 147
B.2 Motivation . . . 148
B.2.1 Le goulot d'étranglement de l'annotation . . . 148
B.2.2 Possibilités de connaissances antérieures et complémentaires . . . 149
B.2.3 Objectif de la recherche . . . 150
B.3 FUN-SIS: une approche entièrement non supervisée pour la segmentation des instruments chirurgicaux . . . 150
B.3.1 PAF-IS: un cadre sans annotation de pixels pour la segmentation d'instances d'outils chirurgicaux . . . 152
B.3.2 KI-BOT: une approche cinématique du goulot d'étranglement pour la régression de la pose d'instruments chirurgicaux flexibles directement à partir d'images . . . 154

B.4 Conclusion 156

List of Figures

1.1 The Great Library of Alexandria in Egypt, the single greatest accumulation of human knowledge in history. While collecting and centralizing information once required incredible effort, nowadays *digitalization* has simplified this process, allowing to accumulate information incredibly faster. Courtesy of [Lib] 23

1.2 Big data analysis is revolutionizing the sport's world, pushing the game to levels never reached before. Courtesy of [Las16, Kid20, McG15]. 23

1.3 Images documenting the first ever laparoscopic cholecystectomy, performed on September 12, 1985, by Erich Mühe. From left to right: model of pistol grip hemoclip applier and scissors used for the procedure; Galloscope-Laparoscope used in the procedure; picture of the abdomen of the patient who underwent the procedure showing portholes in the lower abdomen. Courtesy of [RJ01] 25

1.4 A visual comparison between an OR from early 1900 with a modern-day OR. On the left, a surgeon performing open surgery with minimal equipment, and no information captured about the surgical process. Courtesy of [Jac15]. On the right, the OR of the present, a technological environment teeming with advanced technology and digital information. Courtesy of [Sri21] 26

1.5 Increase in the percentage of procedures performed laparoscopically by surgical residents in a 16-year period for six high-volume interventions: cholecystectomy, inguinal hernia repair, appendectomy, colectomy, gastrectomy, and Nissen. Courtesy of [JCKK20] 27

1.6 Endoscopic camera frame from the GI Genius™ system showing the potential presence of a polyp. The AI prediction, in the form of a bounding-box containing the polyp, is provided to the endoscopist as an overlay on the original frame. Courtesy of [GIG] 29

1.7 Association between surgical instruments and corresponding actions from the CholecT50 dataset for laparoscopic cholecystectomy procedures. Each instrument is mostly used to perform one principal action. 31

1.8 Instrument and trocar usage over time during a cholecystectomy laparoscopic procedure, plotted together with 14 manually annotated surgical phases. Note how each phase transition is associated with the usage of specific combinations of surgical instruments. Courtesy of [PBA+12] 31

1.9 Examples of action triplets from the CholectT50 dataset. Note how, although the liver is present in both frames, it is never part of a triplet. Note also how the spatial proximity of an instrument with an anatomical structure can be used as a hint to identify the target of a triplet. Courtesy of [NYG+22]. 32

1.10 Framework collecting the main CV tasks available to digest endoscopic videos, parametrized by *Time-Scale*, *Spatial-Resolution* and *Semantic Content*. Frame-wise tasks include, from lowest to highest spatial resolution: tool-presence detection, bounding-box (BB) detection, semantic segmentation (binary, tool part, tool type, full-scene), instance segmentation. Short-term tasks include, from lowest to highest semantic content: action/gesture recognition, tool tracking, action triplet and quintuplet detection. Finally, long-term tasks, like step/phase segmentation. 32

1.11 From left to right: original endoscopic frame, binary segmentation, tool part semantic segmentation, tool type semantic segmentation and tool type instance segmentation (separate instances highlighted by their boundary). 37

1.12 Overview of the general framework commonly used to tackle a vision-based 3D tool localisation problem. 39

1.13 Top row shows randomly sampled images from the ImageNet [DDS+09] chair class. Bottom row shows images of different types of laparoscopic scissors, produced by different manufacturers. The scissors, whose appearance is totally constrained by their function, tend to keep clearly recognizable features regardless of the specific type and manufacturer. 42

1.14 Left: shape-priors, in the form of binary segmentation masks of the instruments, obtained from automatic segmentation of chroma-key images. Courtesy of [GPHFD+21]. Right: different views of a 3D CAD model of a laparoscopic forceps head. Courtesy of [Gra] 42

1.15 Left: graphic visualization of instrument coherent motion vs soft tissue incoherent motion using optical flow. Right: laparoscopic triangulation principle for trocar placement (courtesy of [SAG14]), and heat-map representing instrument localisation in the image space during a procedure extracted from the MICCAI 2017 grand-challenge dataset on instrument segmentation [ASK+19] (blue low presence, yellow high presence). Note how the two instruments do not normally overlap and tend to enter the field-of-view from the sides. 43

1.16 Renderings from the LapSim® simulator: from left to right salpingectomy, cholecystectomy, appendectomy and hysterectomy procedures. Courtesy of [Lap20]. 43

1.17 Left: fully-supervised learning framework, where a ground truth signal (GT), usually obtained through manual annotation, is directly used to compute the loss for model training. The model's parameters are optimized to minimize such loss, so that its predictions can match the GT. Right: hypothetical framework using prior/complementary knowledge to generate a pseudo-GT signal for deep learning model training. This framework opens up interesting questions, such as how to generate the pseudo-GT from prior/complementary knowledge and how to effectively learn from it. 44

1.18 Overview of the contributions proposed in this thesis, highlighting the sources of information exploited to replace manual annotations for deep learning model training. Our first contribution (FUN-SIS) tackles the binary segmentation problem, using instrument *shape-priors* and the hypothesis of instrument coherent motion as sources of prior knowledge. Our second contribution (PAF-IS) builds on top of FUN-SIS to solve the instance segmentation task by exclusively relying on general hypotheses on instrument positioning in the field-of-view and binary tool presence labels. Finally, our third contribution (KI-BOT) uses instrument kinematic modelling to learn 3D pose estimation from inaccurate kinematic data only. 45

2.1 Overview of [KJD+21] approach and validation. The approach consists of a standard Mask-RCNN architecture, validated using cross-dataset evaluation with different sampling strategies. Courtesy of [KJD+21]. 49

2.2 Overview of [RZV+18] approach. The approach includes a pre-training step, performed using a pretext re-colorization task on unlabelled data, and fully-supervised training step, on the available labelled data. Courtesy of [RZV+18]. 50

2.3 Images from full-scene laparoscopic simulation (first column) translated into real-looking laparoscopic images (second and third column) using different styles. Courtesy of [ZPIE17]. 51

2.4 Blending process (top row) and blended image sample (bottom picture) created using green-screen recordings of surgical instruments and background-only images. Courtesy of [GPHFD+21]. 51

2.5 Endoscopic images overlaid with corresponding localisation maps and predicted tool centers for different weakly-supervised architectures. Note how the localisation is mostly confined to the tip of the tools. Courtesy of [VMMP18]. 53

2.6 From left to right: endoscopic frame showing the colored marker applied on the laparoscopic instrument; HSV color space histogram of endoscopic frame and colored marker; segmentation masks before and after post-processing spatial filtering. Courtesy of [WAH97]. 53

2.7 Left: Fixed shape template illustration for a suction tube used by [BBO+15]. Right: example of endoscopic frame and segmentation results of boosted decision tree algorithm (green overlay) and template matching (orange overlay). Courtesy of [BBO+15]. 54

2.8 Overview of the 3D pose estimation approach by [AOH+18]. 55

2.9 SSTS [dCRPR19] results on different datasets synthetic and in-vivo datasets. From left to right: original frame; kinematic projection (inaccurate); ground truth segmentation mask (GT); SSTS prediction; fully-supervised model prediction (FSL). Courtesy of [dCRPR19]. 56

3.1 Chapter contribution from the input-output point-of-view. The proposed FUN-SIS approach allows to train a model for surgical tool segmentation requiring as inputs only unlabelled video-clips and tool shape-priors, obtainable in various convenient ways (e.g. by recycling existing annotations from other datasets). The method is based on a novel approach for unsupervised surgical tool segmentation of optical flow images, generating pseudo-label masks, and a newly designed learning-from-noisy-labels strategy, allowing to extract a clean supervision signal to train a single-frame binary segmentation model. 61

3.2 Frames and corresponding optical flow images from DAVIS dataset [PPTM+16] (left) and EndoVis 2017 dataset [ASK+19] (right). Note how, in surgical images, background motion is often coherent and correlated with foreground motion. 62

3.3 General overview of proposed FUN-SIS training architecture. Step I: training of the optical flow segmentation network (*Teacher*, T), as part of a generative-adversarial architecture mapping *shape-priors* into synthetic optical flow images and vice-versa. Step II: *Proxy* segmentation model training, directly supervised by the pseudo-labels obtained from optical flow segmentation by the *Teacher* model. Step III: *Student* segmentation model training on the refined supervision signal obtained by the combination of Proxy predictions and pseudo-labels. Main training losses (L) are also shown. 63

3.4 Overview Step I of FUN-SIS: generative-adversarial training of optical flow segmentation model S^{OF} (*Teacher*), generator (G) and discriminator (D); generated (m^{OF}) and real ($E^{OF}(I_t, I_{t+1})$) optical flow images undergo augmentation via random rotation θ_f. A noise vector $\mathbf{n}$ is concatenated to the shape-prior m to allow one-to-many mapping. Loss boxes (L) are color coded to show which models are responsible for their minimization during training. 64

3.5 Overview of Step II of FUN-SIS: *Proxy* segmentation model training, directly supervised by the pseudo-labels y_t^T, obtained from optical flow segmentation by the *Teacher* model. Loss boxes (L) are color-coded to show which models are responsible for their minimization during training. 66

3.6 Overview of Step III of FUN-SIS: *Student* segmentation model training, leveraging *local* Intersection-over-Union ($IoU^{loc}_{(w,h)}$) between *Teacher* and *Proxy* predictions to select well-labelled regions of y_t^T. $\tilde{L}$ is a pixel-wise loss (e.g. cross-entropy), masked by the pixel-wise multiplication ($*$) with the binarized *local* IoU. Loss boxes (L) are color-coded to show which models are responsible for their minimization during training. 68

3.7 *Local* IoU $IoU^{loc}_{(w,h)}$ is computed by sliding a window of size $w \times h$ on the two input masks, computing standard IoU at each corresponding location. The output is a single-channel image, having the same resolution as the input masks, with each pixel's value being set to the one of the IoU computed for the region it belongs to. 69

3.8 Examples of *shape-priors* used for the EndoVis2017 experiments, and corresponding source image. Top: tools recorded in front of the green-screen and automatically segmented [GPHFD+21], called GrScreenTool; Bottom: frames from multiple robotic-assisted laparoscopic surgeries, manually segmented as part of the RoboTool dataset [GPHFD+21]. Frames (and also masks) in this dataset come with various resolution/aspect ratios. Note how the appearance of the two domains is different: this is mainly due to the fact that GrScreenTool dataset, recorded using an external camera, show a different point of view on the instruments with respect to the standard surgical camera. 72

3.9 Samples from the artificially-corrupted versions of EndoVis2017 dataset. From top to bottom: *Systematic Erosion*, *Erosion & Dilation*, *Tool-Drop*. For each noise source, a sample from D80 (~80% mean IoU between training sample labels and original ones), D60 (~60% mean IoU), D40 (~40% mean IoU), D20 (~20% mean IoU) is shown. 73

3.10 Optical flow segmentation on EndoVis2017VOS. Qualitative results showing frame couples used for optical flow computation, optical flow images after HSV standard conversion, predictions from CIS [YLSS19a] and *Teacher* (trained using RoboTool *shape-priors*), and ground truth (GT). 75

3.11 Surgical tool segmentation on the EndoVis2017VOS dataset. Qualitative results showing, from left to right, input frame I_t, optical flow image y^{OF}_t using HSV standard conversion, predictions from *Teacher* (using RoboTool *shape-priors*), *Proxy*, *Student* and fully-supervised baseline ($\text{Baseline}_{\text{FS}}$), and ground truth (GT). 78

3.12 Surgical tool segmentation on the EndoVis2017VOS dataset: example of boundary masks used for Boundary IoU computation. From left to right, input frame I_t, ground truth mask (GT) and boundary version (B-GT), fully-supervised baseline mask (FS) and boundary version (B-FS), *Student* mask (*Student*) and boundary version (B-*Student*). Boundary IoU is computed as a standard IoU between B-GT and predicted boundary mask (e.g. B-*Student*). 79

3.13 Optical flow estimation inaccuracy leading to missing details in the pseudo-labels (*Teacher* network predictions) such as tool tips. 80

3.14 Box-plots showing IoU distributions from EndoVis2017VOS segmentation experiment (Table 3.2). Fully-supervised baseline $\text{Baseline}_{\text{FS}}$ (grey), *Teacher* (purple 2-step, light purple 3-step), *Proxy* (yellow 2-step, light yellow 3-step), *Student* (blue 2-step, light blue 3-step). 80

3.15 Surgical tool segmentation on the STRAS dataset. Qualitative results showing, from left to right, input frame I_t, optical flow image y^{OF}_t using HSV standard conversion, predictions from *Teacher* (using STRASMasks *shape-priors*), *Proxy* and *Student*, and ground truth (GT). 82

3.16 Qualitative results of the optical flow generator (G), trained using different size of input noise vector among {no-noise,1,32}. First column: input *shape-priors*; first block (x0), no noise concatenation; second block (x1), noise vector of size 1; third block (x32), noise vector of size 32. For each of the 3 blocks, from left to right, the noise vector was smoothly interpolated between all zeros to all ones (trivial for x0, having no concantenated noise). 83

3.17 Analysis of the impact of noise vector size (no-noise, 1, 32) and flow augmentation $AugmFlow$ on optical flow segmentation results by the *Teacher* network on EndoVis2017VOS. Mean IoU [%] is reported. 83

3.18 Analysis of the impact of *Proxy* network's architecture on surgical tool segmentation results of *Proxy* (yellow) and *Student* (blue) networks, on EndoVis2017VOS. Mean IoU [%] is reported. 84

3.19 Analysis of the impact of loss function balancing coefficients α_P and α_S on *Proxy* (yellow) and *Student* (blue) networks, on EndoVis2017VOS. We only consider the case $\alpha_P = \alpha_S = \alpha$; α equal 0 corresponds to cross-entropy loss only, α equal 1 corresponds to log IoU loss only. Mean IoU [%] is reported. . 84

3.20 Impact of local IoU parameters (ϵ_{IoU} and window size w) on effective training size (left) and average effective IoU (right). x-axis can be interpreted as the level of agreement between *Teacher* and *Proxy* required in order to select a certain region (e.g. with ϵ_{IoU} equal to 0.8 a region is considered well-labelled only if the IoU between *Proxy* and *Teacher* predictions for that region is at least 80%). Red markers correspond to $w = 64$ and $\epsilon_{IoU} = 0.5$, the values used in our main experiments. 85

3.21 Analysis of the impact of decreasing *shape-priors* quantity on individual frame and optical flow segmentation, with and without $AugmMask$ augmentation, on EndoVis2017VOS. On the x-axis, the amount of RoboTool *shape-priors* used for training is reported (absolute number and percentage with respect to the total number). Mean IoU [%] for *Student* (blue; dashed: trained without $AugmMask$), *Proxy* (yellow; dashed: trained without $AugmMask$), *Teacher* (purple; dashed: trained without $AugmMask$) is reported. 86

3.22 Analysis of the impact of *unpredictability* and *polarization* noise properties on the proposed method, on the artificially-corrupted EndoVis2017VOS datasets. Top: for each of the 3 noise sources (A, *Systematic-Erosion*, predictable and not-polarized; B, random *Erosion & Dilation*, unpredictable and not-polarized; C *Tool-Drop*, unpredictable and polarized) *Proxy* (yellow) and *Student* (blue) models were trained on the EndoVis2017VOS training dataset, having ground truth labels corrupted with different levels of such noise. The colored bars are meant to improve readability, by visually showing the mean IoU between each training dataset labels and ground truth clean labels (~80% for D80, ~60% for D60, ~40% for D40, ~20% for D20); Bottom: for each set of noisy labels, per-tool IoU histograms ($\text{IoU}_{\text{tools}}$) computed as shown in Figure 3.23, are reported. . 87

3.23 Computation of per-tool IoU between ground truth masks and noisy labels. Left: example of ground truth mask (GT) and noisy label. The smallest region containing each tool in the GT mask is extracted; the same exact region is extracted from the noisy label. Right: Intersection-over-Union (IoU_{tool}) is computed between each region extracted from GT (GT_{tool}) and noisy label (noisy $label_{tool}$); the process is repeated for each tool in each frame of the dataset, and each IoU_{tool} is stored in IoU_{tools}. The distribution of per-tool IoU can then be visualized through histogram plots (Figures 3.22&3.24). 88

3.24 Per-tool IoU histogram (IoU_{tools}), computed as shown in Figure 3.23, for pseudo-labels derived from motion segmentation by the *Teacher* model on EndoVis2017VOS. Note how the distribution tends to be polarized on leftmost bin (completely mislabelled tools) and rightmost bins (*almost*-perfectly segmented tools). 88

3.25 Break-down of the per-tool IoU across the 7 tool classes present in EndoVis2017, for fully-supervised baseline (FS), *Teacher* and *Student* networks. 89

3.26 Surgical tool segmentation by the *Teacher* network in the case of static tools: physiological anatomical movements highlight surgical tools, even while being held still, allowing a proper segmentation by the *Teacher* network. . . 90

3.27 Analysis of proposed method performance when trained on increasing amounts of unlabelled RandSurg data, a dataset consisting of randomly selected surgical videos, downloaded from the public repository [Wor], and tested on EndoVis2017VOS. On the x-axis, the amount of RandSurg frames used for training is reported (absolute number and percentage with respect to the total number). Mean IoU [%] for *Student* (blue), *Proxy* (yellow), *Teacher* (purple) is reported. 91

3.28 Qualitative results on the EndoVis2017VOS dataset, from the experiment reported in Table 3.2. Original frame overlapped with ground truth (blue) and *Student* network's prediction (green). 91

3.29 Qualitative results on the STRAS dataset, from the experiment reported in Table 3.5. Original frame overlapped with ground truth (blue) and *Student* network's prediction (green). 92

3.30 Qualitative results on the Cholec80 dataset. Original frame and overlapping between *Student* network prediction and original frame are shown. Training was carried out using RoboTool *shape-priors*. 92

4.1 Examples of frame-wise and sequence-wise binary tool presence labels for a robot-assisted surgery sequence (each color represents a tool type). All the tools can be attached to the system at the same time, while being visible only in certain frames. 96

4.2 Overview of the proposed Pixel-wise Annotation Free framework for Instance Segmentation (PAF-IS). Top: training architecture highlighting the three core steps. *Tool instantiation* is learnt from binary masks, potentially obtained using recent unsupervised segmentation methods. *Instance-wise feature representation learning* is performed using a contrastive learning strategy, powered by local temporal tracking. This step allows to extract a feature representation of each tool instance in the training set. *Instance-wise tool type classification* is performed by incorporating a minimal amount of human-provided information (*prototype labels*, as few as 8 in our experiments) and cheaply obtainable binary tool presence labels. Bottom: PAF-IS inference architecture. 98

4.3 Overview of the proposed strategy to generate a pseudo-supervision signal to learn instrument instantiation. Given an image I, its binary mask M^B is instantiated using a Connected Component (CC) algorithm, yielding the set of tool masks $\{M_i^{CC}\}$, with i in $[1, N_{CC}]$. From them, the displacement field D^{CC} and the overlap mask M^{OV} can be automatically obtained. A random tool instance is then selected from the training set, and pasted on I, M^B, D^{CC}, producing their augmented versions I^*, M^{B*}, D^{CC*}. 99

4.4 Overview of the proposed instantiation strategy. Given an image I the trained instantiation model predicts the masked displacement field $\tilde{D}$. A square grid is then overlapped to $\tilde{D}$, and the squares with high convergence (per-pixel average $> \varepsilon_C$) are then extracted, and separated by Connected Component (CC) labelling, yielding a set of $\tilde{N}_{Inst}$ centroid regions. Each tool pixel is then assigned to the corresponding centroid, yielding the set of instances masks $\{\tilde{M}_i^{Inst}\}$, with i in $[1, \tilde{N}_{Inst}]$. 100

4.5 Overview of the tracking strategy used to generate positive samples for contrastive learning. Given two consecutive frames, centroids at time t, obtained from the displacement field D_t are mapped to the I_{t+1} space using optical flow OF, computed between the two images I_t and I_{t+1}. The projected centroids are then matched to the ones obtained from the displacement field D_{t+1}. This allows to build the set of tubes $\{T_i\}$, with i in $[1, \tilde{N}_{Inst}]$. Tubes are progressively grown by repeating this process for consecutive frames. 101

4.6 Left: visualization of learnt feature representations of the EndoVis 2017 [ASK+19] training set instances, clustered (N_{km} equal to 8) and projected in the 2D space using t-SNE algorithm [VdMH08]. Each instance point is colored in a different shade of grey to represent the cluster id. Prototype instance features are marked with $\star$, and the corresponding masks are overlaid on the frame and highlighted by a bounding-box, to facilitate their labelling by a human user. The color of the mask overlays represents the ground truth tool type that the user would assign. Right: prototype instance labels propagation to the training set. Each instance-wise feature projection is colored accordingly to its prototype label, assigned via propagation from the prototype instance. 103

4.7 Overview of the proposed weakly-supervised instance classification module. Given an image I, the corresponding set of instance-wise features $\{F_i\}$, with i in $[1, \tilde{N}_{Inst}]$, is obtained from the instance masks $\tilde{M}_i^{Inst}$. Each feature is mapped to the corresponding prototype label S_i^P, which, as shown in this case, does not necessarily correspond to the ground truth label. Each feature is also independently passed through the Teacher (**T**) and Student networks (**S**), yielding the predicted probabilities $\tilde{\mathbf{P}}_\mathbf{i}^\mathbf{T}$, $\tilde{\mathbf{P}}_\mathbf{i}^\mathbf{S}$, and the corresponding predicted labels $\tilde{S}_i^T$, $\tilde{S}_i^S$ (for the sake of readability only the latter are shown in the picture). **T** is trained optimizing the loss L_T computed using the prototype labels $\{S_i^P\}$. Simultaneously, **T** predictions are used to compute the assignment costs $C_{\langle i_C, i_P \rangle}$ for each i_P permutation of each i_C combination of the weak labels $\{S_i^W\}$, with i in $[1, N_W]$. The ordered set $[\bar{S}_i^W]$, with i in $[1, \tilde{N}_{Inst}]$, corresponding to the minimum assignment cost, is used to compute the loss L_S for Student network optimization. 104

4.8 Impact of grid square resolution and threshold value ε_C on the tool instantiation quality for the EndoVis2017 dataset (left) and EndoVis2018 dataset (right). The combination used in our main experiments is highlighted in red. 111

4.9 From left to right, original image, predicted displacement field, and examples of centroid regions and instantiation masks for different combinations of grid resolution and threshold ε_C. Mask colors indicate the ID assigned to the tube the instance belongs to. The combination adopted in our main experiments is highlighted in red. 112

4.10 Left: visualization of the learnt feature representations of the EndoVis 2017 training set instances, projected in the 2D space using t-SNE algorithm [VdMH08]. Each instance point is colored according to the corresponding ground truth tool class. Right: K-Means++ clustering and prototype labels using different number of clusters N_{km}; projected features and prototype instances are colored accordingly to the corresponding prototype labels. . . 112

4.11 Top row: SAM [KMR+23] segmentation results on the EndoVis2017 dataset. Central row: PAF-IS instantiation results obtained from binary annotated masks only. Bottom row: PAF-IS instantiation results obtained from FUN-SIS predicted binary masks only. 114

4.12 Qualitative segmentation results from the EndoVis2017 dataset. Row 1: ground truth; rows 2-5: PAF-IS Student trained on (2) manually annotated binary masks and *frame-wise* tool presence labels, (3) manually annotated binary masks and *sequence-wise* tool presence labels, (4) FUN-SIS predicted binary masks and *frame-wise* tool presence labels, (3) FUN-SIS predicted binary masks and *sequence-wise* tool presence labels. . . 115

4.13 Qualitative segmentation results from the EndoVis2018 dataset. Ground truth and PAF-IS Student results are presented in the same order as Figure 4.12 above. 116

5.1 Application of the general framework for vision-based 3D pose estimation presented in Chapter 1, Figure 1.12. Top: existing approaches commonly learn P_{real} from annotated datasets, and use the distance between the two projections P_{real} and P_{rend} as optimization error to refine the kinematic values $\mathbf{k_s}$, recorded by the robotic system. Bottom left: our solution shifts the target of the optimization to the neural network model Ψ, which directly regresses the kinematic vector $\hat{\mathbf{k}}$ from images. Bottom right: at inference time the predicted kinematics $\hat{\mathbf{k}}$ is obtained from the input image only, by forward propagation through Ψ. 118

5.2 Full KI-BOT training architecture: given an input image I and the corresponding measured kinematics $\mathbf{k_s}$, the training architecture forces a separation between the appearance of the image and its kinematic content. In the bottom branch, the module β transforms I into a *backgroundized* version of itself $\hat{I}_b$. In the top branch, the regressor ψ reduces I to the estimated kinematic configuration $\hat{\mathbf{k}}$. $\hat{\mathbf{k}}$ is then mapped into the binary silhouette projection $\hat{m}$ of the instruments on the image plane by means of the physical model ρ, consisting of the 3D model the instruments r, the robot-camera transformation b and the model of the endoscopic camera c. A decoder ϕ tries to reconstruct the input image I from $\hat{I}_b$ and $\hat{m}$. The model is trained by means of the image-based loss L_r, helped by the auxiliary loss L_a. The *backgroundizer* β and the *physical* module ρ are trained in advance and frozen during the training of regressor and decoder. 120

5.3 *Backgroundizer module* β: the imprecise kinematics $\mathbf{k_s}$ is converted to the binary projection representation m_s, through a robot-renderer model equivalent to ρ. m_s is then expanded to account for uncertainties and used to mask the image I associated with $\mathbf{k_s}$, which is then fed to an inpainting network which produces $\hat{I}_b$, the *backgroundized* version of I. 121

5.4 *Physical* module ρ training: a synthetic dataset is generated by mapping kinematic vectors $\mathbf{k}$ to the corresponding projection masks m, using the 3D kinematic model of the tools, the robot-camera transformation matrix and a potentially non differentiable renderer. The mapping between kinematic vectors and projection masks is directly learnt by the neural network ρ, trained on the synthetic dataset. 122

5.5 STRAS robotic system model, showing the two instruments, each one having 3 joint angles/positions: rotation around the instrument main axis, translation along the same axis, and a cable-actuated bending, here represented as the radius of curvature, but actually measured as the delta between the lengths of the two cables used for actuation. Instrument's main axis forms a small angle ($\sim 10°$) with the main axis of the endoscope (grey cylinder). 124

5.6 Semi-synthetic dataset building: a kinematic configuration **k** and a background b are randomly sampled. **k** is then fed to a VTK model of the robotic instruments ($\rho*$, where * means that binarization of the rendered image is not applied, differently from ρ) and rendered on the image space. The projection is then blended with b to obtain the image I. Parallelly, random noise is added to **k** to simulate the measured kinematics $\mathbf{k_s}$ (visualized as the corresponding binary projection m_s on the image plane obtained through ρ, in order to qualitatively show the difference with the GT kinematic configuration **k**). 125
5.7 Examples of *backgroundizer* module results. Top row shows one image for each dataset (*semi-synthetic*,*phantom*, *in-vivo*), with the corresponding measured kinematics $\mathbf{k_s}$ projection (imprecise). Bottom row shows the corresponding *backgroundized* versions, obtained using the rough localisation provided by $\mathbf{k_s}$. 127
5.8 Examples of qualitative results on the three testing datasets. Top row: *semi-synthetic*; middle row: *phantom*; bottom row: *in-vivo*. The last two columns show corresponding reconstructed image and *backgroundized* image. Note that at inference time none of the two is needed/obtained, since the image-based regressor ψ is a completely independent module. 129
5.9 Left: modelled tool configuration according to a certain kinematic configuration **k**. Right: actual tool configuration due to unmodelled phenomena, like tool slackening and tool-channel interaction. 130
5.10 Potential inference architecture modification, if robot kinematics $\mathbf{k_s}$ is available in real-time. 131
5.11 Potential training architecture modification integrating instance segmentation by PAF-IS to perform instance-wise 3D pose estimation and class-informed decoding. 131
5.12 Top: potential training architecture modification integrating unsupervised tool segmentation by FUN-SIS as P_{real} mapping. The loss L_r could become a cross-entropy between projected and predicted masks. Bottom: application of such architecture to laparoscopic images. Such a solution would need to address the problem of the unknown robot-camera transformation b. 132

6.1 Left: fully-supervised learning framework. Right: general framework adopted in this work for model training with no manually annotated GT. The training follows two paths. Blue: pseudo-GT generation from unlabelled data and prior/complementary problem knowledge. Red: supervision signal denoising, actuated as direct pseudo-GT refinement or loss modulation. Such signal can be obtained from the pseudo-GT signal itself, the unlabelled data and additional prior/complementary knowledge. 134
6.2 Overview of the FUN-SIS approach (Chapter 3) as the application of the general learning framework shown in Figure 6.1 (right). 135
6.3 Overview of the PAF-IS approach for tool instantiation (Chapter 4) as the application of the general learning framework shown in Figure 6.1 (right). . 135
6.4 Overview of the PAF-IS approach for instance classification (Chapter 4) as the application of the general learning framework shown in Figure 6.1 (right).136

6.5 Overview of the KI-BOT approach for 3D pose estimation (Chapter 5) as the application of the general learning framework shown in Figure 6.1 (right). . 136

B.1 Évolution de la salle d'opération. 147
B.2 Aperçu des principales contributions de la thèse. 150
B.3 Vue d'ensemble de l'architecture FUN-SIS. 151
B.4 Résultats de FUN-SIS sur l'ensemble de données EndoVis 2017, et comparaison avec la ligne de base entièrement supervisée. 153
B.5 Vue d'ensemble de l'architecture PAF-IS. 153
B.6 Vue d'ensemble de l'architecture KI-BOT. 155
B.7 Résultats qualitatifs sur l'ensemble de données STRAS in-vivo. 156

List of Tables

1.1 The table reports an overview of the main surgical instrument segmentation tasks, organized according to their formalization in Semantic, Instance and Pan-Optic (Pan.). Each task is characterized by its semantic scope. The main datasets annotated for each task are also reported. Datasets included are the segmentation datasets of the MICCAI EndoVis challenges 2015, 2017 [ASK+19], 2018 [AKB+20], 2019 [RRF+20], CaDIS [GFK+19], HeiSurf; CholecSeg8k [HKK+20]; Endoscapes [AMV+21]. 36

2.1 Methods presented in Section 2.3, and thesis contributions (FUN-SIS, PAF-IS, KI-BOT, last three rows), exploiting prior and complementary information to solve various instrument localisation tasks. For each solution we highlight, from left to right: the task, among binary segmentation (BS), instance segmentation (IS), bounding-box detection (BB) and 3D pose estimation (3D); if Deep Learning based (DL) or not; if based on simulation/semi-synthetic data (Simul.), complementary information (Compl.), in the form of kinematics (K) or binary tool presence information (P), weak or strong prior knowledge (PK); if requiring manual annotations, in the form of segmentation masks (Segm.), at some step of the method. 57

3.1 Optical flow segmentation. Comparison of the proposed method (*Teacher*), using different *shape-priors* for training (RoboTool, GrScreenTool for EndoVis2017VOS experiments; FBMS, SegTrackV2 for DAVIS2016 experiments), with the state-of-the-art CIS approach (without and with post-processing, in parenthesis, taken from [YLL+21]) and a fully-supervised baseline ($\text{Baseline}_{\text{FS}}$). Mean IoU [%] and standard deviation are reported. The percentage of annotated training samples required by each method is also reported (Annot. [%]). Note that CIS (*) uses frames and optical flow to make predictions, while our approach only uses optical flow. 74

3.2 Surgical tool segmentation of individual frames. Comparison of the proposed unsupervised method (trained using RoboTool *shape-priors*), with state-of-the-art unsupervised AGSD [LWJ+20a] approach, fully-supervised approaches TernausNet-16 [SRKI18a] and MF-TAPNet [JCDH19a], and fully-supervised baseline ($\text{Baseline}_{\text{FS}}$) on the EndoVis2017VOS and EndoVis2017Challenge datasets. Results in parenthesis for state-of-the-art approaches were obtained by training the models using the code released by the authors. Mean IoU [%] and standard deviation are reported. The percentage of annotated training samples required by each method is also reported (Annot. [%]). Note that MF-TAPNet uses 2 consecutive frames at inference time to make a prediction, while the other approaches use individual frames. 77

3.3 Statistical analysis of tool segmentation results obtained in EndoVis2017VOS (Table 3.2). For each pair, t-test was run (*p-values* reported in first column) and *Cohen's d* number was computed. 77

3.4 Surgical tool segmentation of individual frames. Results of the proposed method on the EndoVis2017VOS dataset using RoboTool *shape-priors*. Mean Boundary-IoU (B-IoU) [%] and standard deviation are reported. The percentage of annotated training samples required by each method is also reported (Annot. [%]). 78

3.5 Surgical tool segmentation of individual frames. Results of the proposed method on the STRAS dataset using STRASMasks *shape-priors*. Mean IoU [%] and standard deviation are reported. The percentage of annotated training samples required by each method is also reported (Annot. [%]). . . . 81

3.6 Analysis of the impact of the *shape-priors* dataset on frame segmentation. Comparison of the proposed method trained using RoboTool and GrScreenTool as *shape-priors* on EndoVis2017VOS. Mean IoU [%] and standard deviation are reported. 85

3.7 Per-tool segmentation results on the EndoVis2017VOS dataset separated by surgical tool class: *Bipolar Forceps* (BF), *Prograsp Forceps* (PF), *Needle Driver* (ND), *Vessel Sealer* (VS), *Grasper Retractor* (GR), *Scissors* (S), *Other* (O). Mean IoU [%] is reported for fully-supervised baseline ($\text{Baseline}_{\text{FS}}$), *Teacher* and *Student* networks. 89

4.1 Tool instantiation results for the proposed PAF-IS approach and Mask-RCNN on EndoVis 2017 and 2018 datasets, trained according to three modalities: fully-supervised (GT) and unsupervised using Connected Component labelling of manually annotated masks (CC_{M}) and FUN-SIS predicted masks (CC_{F}). 108

4.2 Instance segmentation results for the proposed PAF-IS approach, state-of-the-art methods on EndoVis 2017 and 2018 datasets. Supervision signals used by each approach are reported: pixel-wise semantic labels (S), pixel-wise instance labels (I), required by fully-supervised instance segmentation approaches, pixel-wise binary segmentation masks (B, for PAF-IS, if not checked FUN-SIS predicted masks are used), prototype labels (P, 8 labels in total in these experiments, ~0.3% of total training instances), frame-wise tool presence labels (FW) and sequence-wise tool presence labels (SW). † methods using temporal information at inference time. ‡ methods using additional tool-part annotations for training. 109
4.3 Results of the ablation study on unsupervised instrument instantiation from manually annotated binary masks, highlighting the separate and combined impact of: masking of potentially overlapping instances (OV) and pasting of random tool instances (PS). 110
4.4 Results of the ablation study on unsupervised instrument instantiation from FUN-SIS predicted binary masks, highlighting the separate and combined impact of: masking of potentially overlapping instances (OV) and pasting of random tool instances (PS). 110
4.5 Results of the ablation study investigating the impact of the number of clusters N_{km} on final segmentation results, evaluated using challenge IoU metric. Results obtained using a): manually annotated binary masks on the EndoVis2017 dataset, b): FUN-SIS predicted binary masks on the EndoVis2017 dataset, c): manually annotated binary masks on the EndoVis2018 dataset, d): FUN-SIS predicted binary masks on the EndoVis2018 dataset. Best result across the number of clusters highlighted in bold. 113

5.1 Number of images in each training and testing dataset (* GT 2-D segmentation mask only, obtained via manual segmentation) 126
5.2 Semi-synthetic dataset results. Comparison with raw kinematics $\mathbf{k_s}$ and fully supervised methods BSupK$_s$, BSupSoftK$_s$. For the joint mean absolute errors (translation: *tr.*, rotation: *ro.*, bending: *be.*), lower is better. For the reprojection IoU metric higher is better. 128
5.3 Evaluation of the IoU on the real datasets (Phantom & in-vivo). Comparison with raw kinematics $\mathbf{k_s}$ and fully supervised methods BSupK$_s$, BSupSoftK$_s$. For the in-vivo the average of the results obtained for each day is reported. . 128

6.1 Comparison between pseudo-GT quality and final model's performance, after training on the denoised signal. Results for our contributions on the tasks of binary segmentation (BS), tool instantiation (TI), instance classification (IC) and 3D pose estimation (3D). The reported metrics were taken from the main results tables in Chapters 3, 4 & 5. 137

List of Abbreviations

AI	Artificial Intelligence
BS	Binary Segmentation
CC	Connected Components
CNN	Convolutional Neural Network
CV	Computer Vision
DL	Deep Learning
GAN	Generative Adversarial Network
GT	Ground Truth
IoU	Intersection over Union
IS	Instance Segmentation
OR	Operating Room
SeS	Semantic Segmentation
SDS	Surgical Data Science
VOS	Video Object Segmentation

Glossary of Notation

b	Robot-camera transformation matrix
c	Endoscopic camera model
L_{CE}	Cross-entropy loss
L_{SCL}	Supervised contrastive loss
m	Shape-prior mask
N_{cls}	Total number of tool classes
r	3D kinematic tool model
S^W	Binary tool presence label
S^P	Prototype label

CHAPTER 1

Introduction

Contents

1.1 Background 22
1.1.1 The Big Data Gold Rush 22
1.1.2 Healthcare in the Big Data Era 23
1.2 Surgery in the Big Data Era 24
1.2.1 The *datafied* Operating Room in the Minimally Invasive Surgery Era 25
1.2.2 Emerging Challenges in Minimally Invasive Surgery 27
1.2.3 Opportunities for Surgical Data Science 28
1.2.4 The Value of Endoscopic Video Analysis 29
1.3 Surgical Computer Vision 30
1.3.1 *Instrument-Centrality* of Surgical Computer Vision tasks 30
1.3.2 From Building-Blocks to Applications 33
1.4 Surgical Instrument Localisation and Identification 34
1.4.1 Problem Statement 35
1.4.2 The Annotation Bottleneck 39
1.4.3 The Potential of Prior and Complementary Knowledge 41
1.4.4 Research Question 43
1.5 Thesis Contribution 44
1.6 Outline 46

1.1 Background

According to the World Economic Forum, we are currently living on the brink of the fourth industrial revolution [Sch16]. During the first industrial revolution steam power allowed to mechanize production. Electric power enabled mass production during the second revolution, and the use of electronics, computers, and the Internet promoted digitization during the third one, shaping the world we are currently living in. More importantly, this digital revolution created the perfect substrate for the next and fourth revolution, which is bound to definitely blur the boundaries between the physical and digital worlds. This transformation is already underway, and the pace at which this is happening has no historical precedent. Artificial Intelligence (AI) has quickly become part of our everyday lives through a wide range of applications such as chatbots, navigation apps, writing assistants, and smart web search engines, with many more ready to be integrated into our society. The impressive progress made by AI in recent years has been enabled by a combination of hardware capability improvements and algorithmic breakthroughs, and fueled by a historically unprecedented availability of data. Indeed it can be argued that the boundaries between the physical and digital worlds have started fading long before AI became part of our lives: as soon as digital technologies were introduced, several aspects of our reality suddenly became describable by means of *digital data*. Electronic payment systems, electronic mails, social networks, and electronic health records are examples of technologies that made important aspects of our lives - like social interactions, buying habits, and health monitoring - describable by means of digital data. In the late '90s the term *big data* was coined to describe the astonishing amount of heterogeneous data produced, at high velocity, by digital technologies [Pre13]. The possibility to quantify complex problems through big data is widely regarded as the lifeblood of the fourth industrial revolution [Wel19].
The next sections introduce this digital data revolution, specifically focusing on how it is bound to impact the way high-quality healthcare can be delivered to patients.

1.1.1 The Big Data Gold Rush

The amount of digital data produced worldwide is doubling every two years [Loh12], and this growth rate is increasing. In the early 2000s, only one-quarter of all the world's stored information was in digital form. Driven by the explosion of the internet, the availability of cheaper sensors integrated into everyday objects like cellphones, the internet-of-things, and more digital technologies, digital data have grown to represent more than 96% of the total information globally stored [CMS13]. In absolute numbers, the amount of digital data globally stored is projected to reach 181 zettabytes in 2025. To put it into perspective, in 2020 enough digital information was produced worldwide to give every person alive more than 300 times the amount of information that was stored in Alexandria's Library, once believed to house the sum of the entire human knowledge (Fig. 1.1). Given these figures, one should not be misled into thinking that big data today are still just a byproduct of the digital technology we consume. In a 2013 article [CMS13], Kenneth Cukier, senior editor of The Economist, introduced the term *datafication* to define the way big data were going to impact our society. *Datafication* was described as the idea of turning every aspect of our lives into digital data: not as a mere digitalization of existing analog data, but as a quantification and modelling of the whole reality surrounding us.

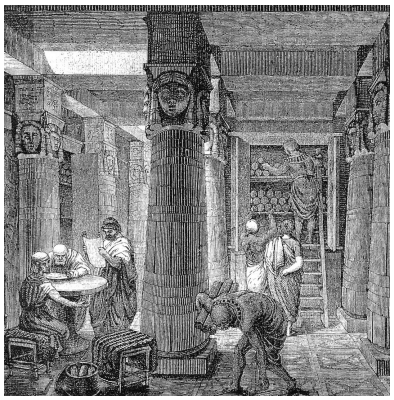

Figure 1.1: The Great Library of Alexandria in Egypt, the single greatest accumulation of human knowledge in history. While collecting and centralizing information once required incredible effort, nowadays digitalization has simplified this process, allowing to accumulate information incredibly faster. Courtesy of [Lib]

Today, only ten years later, big data have already permeated many aspects of our lives. Businesses are using them in a great variety of ways, specialized and tailored to individual needs. Big data analysis is used, for example, in finance to study and anticipate the stock market, exploiting the *datafication* of heterogeneous factors like social trends, economic factors, and political landscapes. Transportation companies use data to improve driving behaviour, optimize routes, and anticipate vehicle maintenance. Professional sport teams work in close contact with data analysts to optimize recruiting [Bea22]; coaches use data analysis to prepare customized game plans based on opposing teams, and to manage athletes' workload to maximize long-term results; even individual athletes have started using data analysis to quantify their value to their teams and to negotiate contracts [Pri21]. Data analysis is transforming professional sport from an artisanal craft based on the individual experience of recruiters, coaches, and athletes, into an evidence-based science built on objective data analysis (Fig. 1.2).
Beyond businesses, the possibility to *quantify*, *predict* and *optimize* problems offers unique opportunities in the healthcare sector, which inevitably found itself on the edge of this revolution.

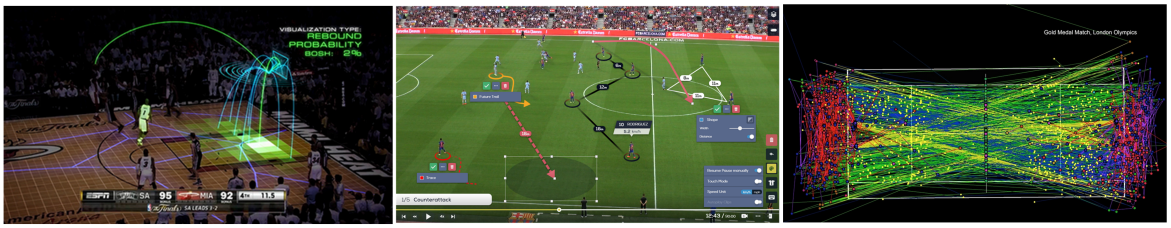

Figure 1.2: Big data analysis is revolutionizing the sport's world, pushing the game to levels never reached before. Courtesy of [Las16, Kid20, McG15].

1.1.2 Healthcare in the Big Data Era

In the current landscape of a constantly expanding and ageing population [CDRV09], health services are required to become more efficient, accessible, and sustainable.
Data science, the interdisciplinary field aimed at analyzing data, holds the potential

to improve the quality of patient care, by enhancing our understanding of complex problems, from disease development to epidemic spreading dynamics, predicting their evolution, and delivering optimized and anticipated solutions. As the process of *datafication* is turning reality into data, statistical analysis can be used to reveal meaningful hidden patterns in them, and Artificial Intelligence (AI) can mine this knowledge to optimize the present and anticipate the future.

Translational success stories in healthcare already exist. As an example, data science and AI have been successfully used in 2015 during the West African Ebola virus outbreak, to optimize and speed up the screening of compounds capable of binding to a glycoprotein that could prevent Ebola virus penetration into cells. This analysis, which typically would have taken months or years, was completed in less than a day [Mes17]. Beyond drug discovery, data science and AI have been proven able to enhance clinicians' work. For example, the practice of colonoscopy is being revolutionized by the use of AI for computer-aided polyp detection, already supported by strong clinical evidence [HSI+21]. In the field of radiology, commercially available products like S-Detect (Samsung Medison, Co., Ltd., Seoul, Korea) can already be used to support operators in the interpretation of ultra-sound images for early-stage diagnosis of cancer [ZJY+21]. Together with clinical practice, AI is also impacting hospital management: Johns Hopkins Hospital, for example, implemented a program to optimize the efficiency of patient operational flow using predictive AI, with a drastic improvement in its ability to quickly admit and discharge patients [For19].

The impact of big data and AI on healthcare is prominent and horizontal, and testified by tangible changes in health services from which the population is already benefiting. Strikingly, such success stories are still lacking in surgery, a critical segment of healthcare. The following paragraphs explore the present state of big data in surgery, with a specific emphasis on the intra-operative phase, the core of surgical care, and the primary focus of this thesis.

1.2 Surgery in the Big Data Era

With approximately 30% of the entire global burden of disease requiring surgical management and over 330 million procedures performed annually, surgery represents a critical segment of healthcare systems worldwide [SBAM15, WHM+15]. Reports from 2010 suggest that inpatient surgical care accounts for nearly 50% of all hospital expenditures and 30% of overall healthcare costs [MMIW10]. Today, only ten years later, these figures are widely regarded as an underestimation, as they do not account for post-operative care following inpatient surgery, re-admissions, and outpatient elective surgery [KLO+20]. More importantly, regardless of the significant improvements to surgical techniques that occurred over the last few decades, surgical care remains perilous, variable, and opaque: post-operative deaths still account for 8% of all deaths globally, making it the third greatest contributor worldwide [NMB+19]. The criticality of surgery for healthcare is also evident when assessing its impact on patient care through Adverse Events (AEs) analysis. An AE is defined as *"an unintended injury or complication resulting in a prolonged length of hospital stay, disability at the time of discharge or death caused by healthcare management and not by the patient underlying*

disease" [BNF+04]. From a structured record review study on 7,926 patients carried out in Dutch hospitals in 2011 it emerged that surgical AEs occurred in 3.6% of hospital admissions and represented 65% of all AEs. Among AEs, 65% involved human factors as root causes, and 41% were considered preventable [ZdBdK+11]. Disconcerting statistics and immeasurable human cost aside, intra-operative care remains an extremely siloed segment of patient care that is both insufficiently analyzed and poorly documented. To date, the standard approach to the documentation of surgical intervention is the generation of narrative operator reports that have been proven to be inadequate, unreliable, and inherently subjective across different types of surgical interventions [WST+13, SHW+10].

Pushed by such long-standing challenges, surgical technique has kept evolving over the last decades. The irruption of digital technologies into the surgical practice has redesigned the Operating Room (OR), and the figure of the surgeon, which had to quickly adapt to new ways of doing surgery. As for many fields outside healthcare, this digital revolution has laid the foundation for the integration of big data into surgical practice [TBB17]. The following paragraphs describe the present OR - not long ago *"the operating room of the future"* [CKM05] - from a *datafication* perspective, exploring the challenges and opportunities which ultimately motivate and enable and the work presented in this thesis.

1.2.1 The *datafied* Operating Room in the Minimally Invasive Surgery Era

On September 12th, 1985, Dr. Erich Mühe of Böblingen, Germany, performed the first ever laparoscopic cholecystectomy [RJ01]. During this procedure, a patient's gallbladder was successfully removed without any large incision of the abdomen. Two small incisions (<2cm) were performed in the lower abdomen to introduce instruments (grasper, pistol grip hemoclip applier, and scissors) in the patient cavity. An endoscope was introduced through the umbilicus into the peritoneal cavity to visualize the surgical site (Fig. 1.3).

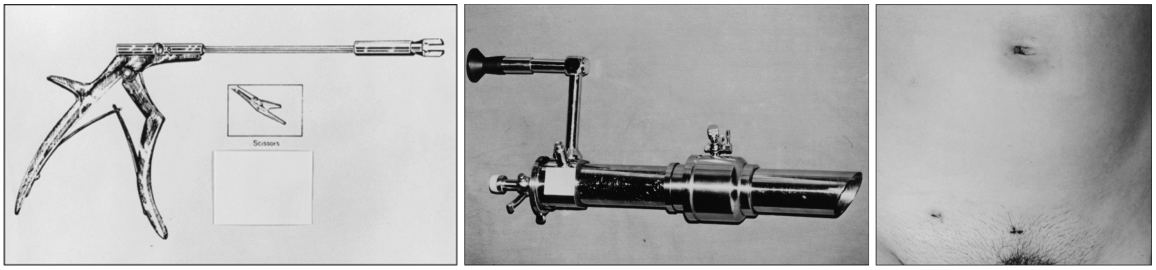

Figure 1.3: Images documenting the first ever laparoscopic cholecystectomy, performed on September 12, 1985, by Erich Mühe. From left to right: model of pistol grip hemoclip applier and scissors used for the procedure; Galloscope-Laparoscope used in the procedure; picture of the abdomen of the patient who underwent the procedure showing portholes in the lower abdomen. Courtesy of [RJ01]

This first laparoscopic cholecystectomy represented the culmination of a decades-long series of technical innovations, like the introduction of fiber optics and Hopkins rod-

lens, and of successful applications of the laparoscopic techniques to increasingly complex problems, from biopsy collection in the early '40s to gynecological procedures and appendectomies in the late '70s [Nak17]. Furthermore, this event promoted the use of laparoscopy for other surgical procedures, such as hysterectomy and nephrectomy, performed laparoscopically for the first time in 1989 and 1990, respectively [RDM89, CK16], and the consolidation of other minimally invasive surgical techniques, like flexible endoscopic surgery [MRD+07].
In 2003, the American surgeon Richard Satava, described as "disruptive" the change that the first laparoscopic cholecystectomy represented for the whole surgical world, as it marked the transition from the *"Industrial Age"* of surgery (of which laparoscopy was still a product) to the *"Information Age"* [Sat03]. The real revolution extended beyond the direct advantages coming from the minimal invasivity, lying in the fact that endoscopic surgery was a transition technology to computer-enhanced and image-guided surgery.
In order to enable this new kind of surgery, the OR was forced to evolve, advancing towards the highly technological environment it is today (Fig. 1.4).

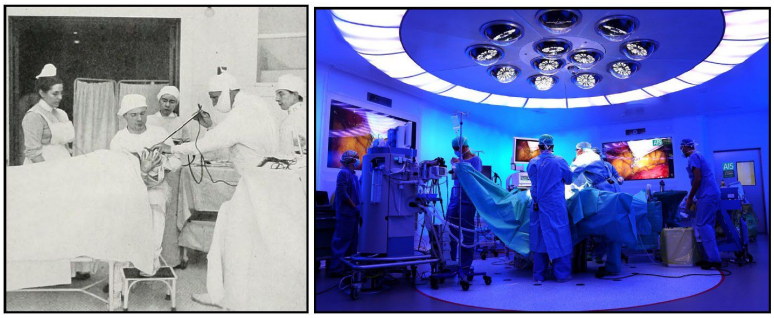

Figure 1.4: A visual comparison between an OR from early 1900 with a modern-day OR. On the left, a surgeon performing open surgery with minimal equipment, and no information captured about the surgical process. Courtesy of [Jac15]. On the right, the OR of the present, a technological environment teeming with advanced technology and digital information. Courtesy of [Sri21]

ORs are currently populated with digital devices, potentially able to capture the totality of the intra-operative phase through a variety of signals: activation signals from electrical instruments, vital signs of the patient, signals from anesthesia machine, videos from endoscopic cameras, CO2 pressure values; robotic systems such as the da Vinci® allow to capture additional signals, like endoscopic stereo-camera videos, kinematic joint values, surgical instrument usage, logs from master's console. Furthermore, this environment is fertile for the introduction of additional sensors, such as ceiling-mounted cameras and audio recording systems.

Data produced in the OR can be characterized by the original 3 V's of big data: *volume*, *velocity* and *variety*. First of all, they are produced, and potentially stored, at high velocity *velocity* and in large *volume*: if limited to laparoscopy, a striking 13 million procedures are currently performed each year, based on recent market studies [iR20]. The number is inevitably bound to increase: among the 300+ million estimated total surgical procedures performed globally each year (both open and minimally invasive) only ~6% occur in the poorest countries, where over a third of the world's population lives [Ric16]. As the development of safe, essential, life-saving surgical and anaesthesia care in low-income countries is considered a global priority, the absolute number of surgeries is inevitably bound to increase in the future. Furthermore, in the last 20 years the laparo-

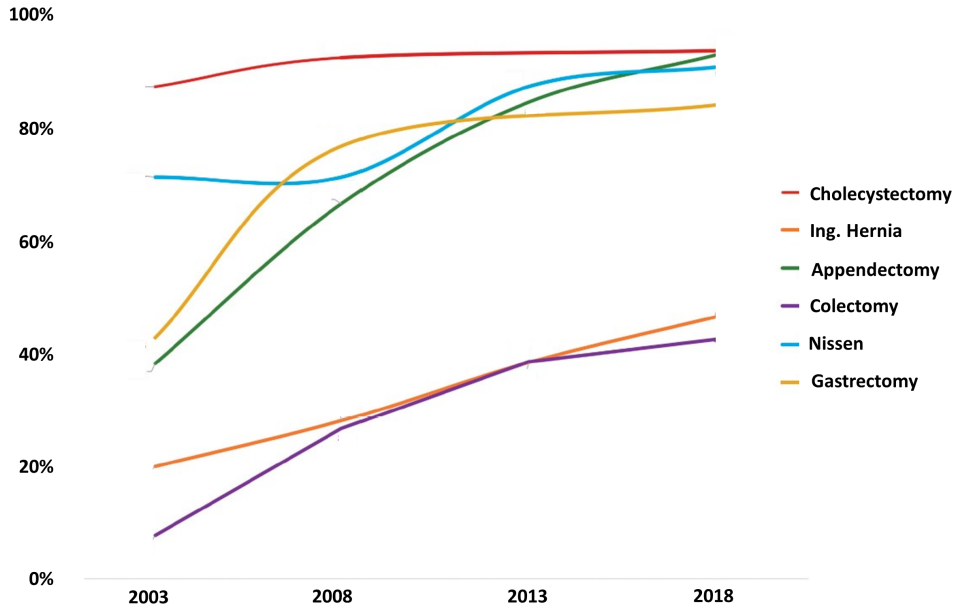

Figure 1.5: Increase in the percentage of procedures performed laparoscopically by surgical residents in a 16-year period for six high-volume interventions: cholecystectomy, inguinal hernia repair, appendectomy, colectomy, gastrectomy, and Nissen. Courtesy of [JCKK20]

scopic approach has been proven to be preferable to open surgery for several procedures such as pancreatic and hepatic resections [CMB+18], cholecystectomy [AAK+14], appendectomy [BDSF+16] and inguinal hernia [TKM+20]. The reasons behind this reside in the small incisions required to perform such procedures laparoscopically, reducing risks and discomforts for the patient (lower infection rate, minimal post-operative pain), and the hospitalization time. This evidence has resulted in a constant increase in the percentage of procedures performed laparoscopically [JCKK20] (Fig. 1.5). Analogously to laparoscopy, other minimally invasive techniques, like flexible endoscopic intervention, are progressively increasing in volume [MB18]. All these concurring factors are indirectly increasing the availability of digital information capturing intra-operative care.
In addition to *volume* and *velocity*, OR data are also characterized by a great *variety*. Due to the diversity of digital technologies housed in the OR, the data produced are extremely heterogeneous, and in some cases redundant (i.e. the same phenomenon is described by multiple data sources). Main data types include 1D time signals (electronic tools activation, vital signs, anesthesia signals, kinematic joint values), image data (videos from endoscopic cameras, ceiling cameras, intra-operative imaging systems), text (handwritten or digital surgical reports, patient's pre-operative data) and, potentially, audio recordings [JJLG20].

1.2.2 Emerging Challenges in Minimally Invasive Surgery

While improving clinical outcomes for several interventions, minimally invasive surgery has significantly increased the cognitive burden on surgeons [ZMSO20]. Such mental workload is necessary to compensate for the reduced instrument dexterity, the complex hand-eye coordination, and the lack of 3D perception, compared to open surgery. These factors place great mental stress on surgeons, commonly associated with an in-

creased probability of adverse events, as well as with a negative impact on team dynamics [ZLZ+19]. Additionally, the technical skills required to perform minimally invasive surgery are hard to master [HCS+14] and have required setting up simulation training programs like the Fundamentals of Laparoscopic Surgery (FLS) program [SFV+10]. The resulting prolonged learning curve amplifies the gap in surgical skills between experienced surgeons and residents, making lack of experience a factor highly associated with surgical errors [SHK+20]. While these factors can indirectly increase the chances of adverse events, misjudgement due to the lack of 3D perception in minimally invasive surgery can often be a direct cause of severe surgical errors. In cholecystectomy, for example, the laparoscopic approach has mostly replaced the open one, supported by evidence of reduced complications and shorter hospital stay [AAK+14]. Nonetheless, the rate of a severe complication, known as the Bile Duct Injury (BDI), drastically increased 2-4 fold with the advent of laparoscopy [SC00]. It is estimated that over 97% of BDIs result from a visual illusion due to lack of 3D perception, leading to the misidentification of patient's anatomy [WSG+03]. Cases of optical illusion have been also documented for different procedures beyond cholecystectomy [Ans18].

However, surgical challenges are now more than ever observable. The introduction of digital systems has opened a window in the OR: beyond observable, surgery can now be *datafied*, opening opportunities for *quantification*, *anticipation* and *optimization* of surgical practice. These new possibilities have led to the formalization of the dedicated interdisciplinary research field of surgical data science.

1.2.3 Opportunities for Surgical Data Science

Surgical Data Science (SDS) is an emerging research field born with the aim to *"improve the quality of interventional healthcare through the capture, organization, analysis and modelling of data"* [MHVS+17].
In the present OR, almost every factor potentially impacting surgical outcomes can be documented and quantified. OR staff interactions, auditory and cognitive distracting factors, surgical workflow, surgical performance, incidence and severity of complications are just some of the aspects that can be objectively documented by data. Through quantification, such aspects can be modelled and linked to surgical outcomes for improved understanding and subsequent optimization of surgical practice. As an example, initial work on data systematically recorded in the OR, including video-audio data from ceiling cameras, microphones and endoscopic cameras, has been used to investigate aspects like auditory and cognitive distractions incidence in the OR, and their correlation with surgery duration, surgeon performance and surgical events, extracted from endoscopic videos [JJLG20]. The surgical act can then be improved by providing real-time enhanced information to the surgeon to support its decisions: this includes timely notifications of upcoming critical steps, dynamic surgical checklists, augmented reality and more. Surgical training can be sped up by systematic performance assessment, as well as by targeted reviewing focused only on critical events, maximizing the ratio between the information conveyed and the time required for reviewing. Optimization can be applied to OR management as well, for example through remaining surgical duration estimation enabling early patient preparation and streamlined procedures.

Among the data available in the OR, endoscopic videos represent a particularly valuable and versatile source of information to model the surgical process and enable SDS

applications. The next section explores their value for SDS, highlighting the need for tools to automatically process them.

1.2.4 The Value of Endoscopic Video Analysis

Video data from endoscopic cameras are the piece of online information on which surgeons rely the most to take critical decisions during minimally invasive surgical procedures. Several factors contribute to make endoscopic video analysis an essential piece of SDS:

- Endoscopic videos provide *comprehensive* information about patient anatomy, surgical instrumentation in use during a procedure, and interactions between the two. Endoscopic videos allow to study the surgical workflow, to spot deviations from the standard one, as well as the reasons behind deviations, to identify adverse events [FCC+18], critical events and potentially critical events like near-misses [BGG15] and to assess the skill level of surgeons [CFM+20].

- Endoscopic videos capture an *extensive* portion of intra-operative patient's care, coinciding with anatomical manipulation by the surgeon. While phases like patient preparation, anesthesia and trocar placement are excluded and would require different sensory data to be assessed, the portion captured is often regarded as the most critical and the least documented one [ZdBdK+11, WST+13].

- Endoscopic videos coincide with the surgeon's *point-of-view* of surgery. The video feed captured by the endoscopic camera is displayed on multiple monitors inside the OR and directly used by the surgeon as main sensory feedback. This has two crucial implications: endoscopic videos can be analyzed to investigate *how* surgeons took certain decisions based on the exact information they receive. As an example, in laparoscopic cholecystectomy, the visual perception illusion resulting in bile duct injuries can be clearly documented only from endoscopic videos [SW07]. Secondly, endoscopic videos can be directly used as a vector to provide enhanced information to the surgeon. This is already in place outside the surgical world for AI-assisted colonoscopy, where the information about AI-detected polyps is provided to the clinician through overlays on the video feed streamed on the monitor (Fig. 1.6).

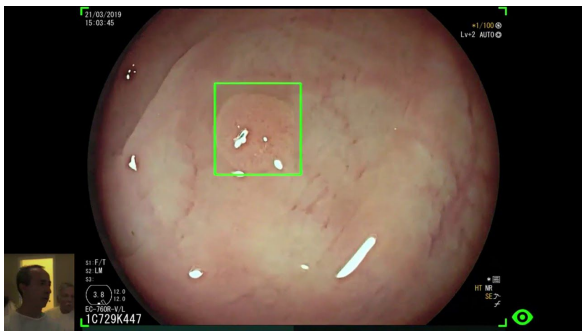

Figure 1.6: Endoscopic camera frame from the GI Genius™ system showing the potential presence of a polyp. The AI prediction, in the form of a bounding-box containing the polyp, is provided to the endoscopist as an overlay on the original frame. Courtesy of [GIG]

1.3 Surgical Computer Vision

The features described above make endoscopic videos the ideal source of information to quantify and study the surgical process (through workflow analysis, surgical skill evaluation, adverse event detection, and post-operative documentation) and, potentially, to extract the necessary context to promptly provide intra-operative support to the surgeon (through notifications, checklists, augmented information). However, the unstructured nature of raw endoscopic videos represents a barrier to extensive storage and usage of such information. Hospitals do not currently own the infrastructures necessary to store such volumes of raw data; in addition, the lengthiness and complexity of raw videos prevent their use for post-operative video reviewing and skill-assessment, as their interpretation would require a significant human effort.

Computer Vision (CV) is a sub-branch of AI focused on building algorithms and methods for understanding the information captured in images and video. To make them tractable, vision problems are broken down into minimal building blocks, such as object identification, localisation, action/activity recognition [CVMS20]. In the surgical context such CV tasks are tailored to capture relevant information about the surgical workflow. Relevant tasks used to extrapolate tractable information from endoscopic videos are, for example, instrument presence detection, instrument localisation, anatomy identification, phase/step segmentation, gesture recognition, tool-tissue interaction estimation.

In the next paragraphs we introduce such surgical CV tasks, and discuss their use in the context of applications that can directly impact surgical care. In particular, Section 1.3.1 highlights the concept of *instrument-centrality* of most surgical CV tasks, introduced in [NYG+22] for the task of tool-tissue interaction estimation. *Instrument-centrality* defines the drive that surgical instrument identification and localisation have on other surgical CV tasks. This concept lays the foundations for the work presented in this thesis, focusing on CV methods for instrument localisation and identification from endoscopic videos.

1.3.1 *Instrument-Centrality* of Surgical Computer Vision tasks

Surgery involves the *"manipulation of a target anatomical structure to achieve a specified clinical objective during patient care"* [MHEF+18]. During surgery, patient anatomy gets manipulated by the surgeon through the surgical instruments: the way instruments interact with patient anatomy is the final result of surgeon's cognitive process and technical ability, and directly contributes to determine the surgical outcome.
Surgical instruments have evolved to effectively carry out specific tasks like *grasping*, *dissecting*, *cutting*, *clipping* and *suturing* tissue. For example, in laparoscopic cholecystectomy procedures, a total of six different surgical instruments are commonly used: *bipolar forceps*, *clipper*, *grasper*, *hook*, *irrigator* and *scissors*. Each of them has a very specific use as can be observed by statistics on the public CholecT50 dataset [NP22] (Fig. 1.7). In this sense *action recognition* in surgical CV is a highly *instrument-centric* task, where tool type recognition can guide action recognition.

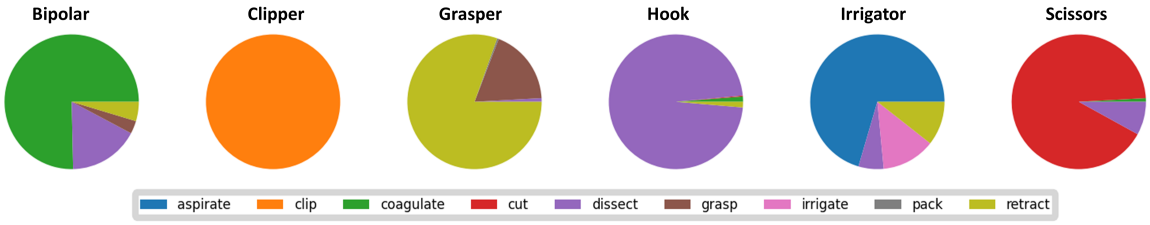

Figure 1.7: Association between surgical instruments and corresponding actions from the CholecT50 dataset for laparoscopic cholecystectomy procedures. Each instrument is mostly used to perform one principal action.

Even when reducing the temporal granularity of activity recognition to surgical steps or phases, the sub-tasks which constitute a surgical procedure [GKL+21], the correlation between steps/phases and instrument usage remains clear. In early surgical workflow analysis works for example, instrument usage was used to infer surgical phases in laparoscopic cholecystectomy procedures [PBA+12] (Fig. 1.8).

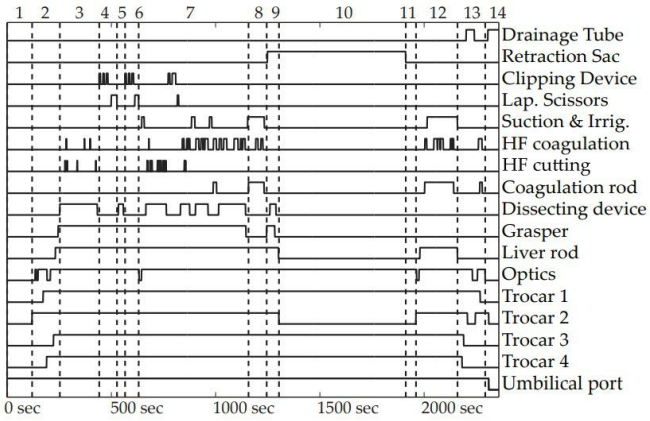

Figure 1.8: Instrument and trocar usage over time during a cholecystectomy laparoscopic procedure, plotted together with 14 manually annotated surgical phases. Note how each phase transition is associated with the usage of specific combinations of surgical instruments. Courtesy of [PBA+ 12]

Instrument-centrality can be even better appreciated when considering the *target* definition in the task of tool-tissue interaction estimation, also known as *action triplet recognition*. This task aims at jointly identifying instrument type, action carried out and subject of such action (defined as *target*, often an anatomical structure). In this task, visibility alone does not determine the consideration of a certain anatomical structure as part of a triplet. Instead, the identification of the target is driven by its interaction with a tool. To a first approximation, the relative position of surgical instruments and anatomical structures can be an informative cue to identify the target of that instrument's action, as in most cases surgical actions involve direct contact between the two (Fig. 1.9).

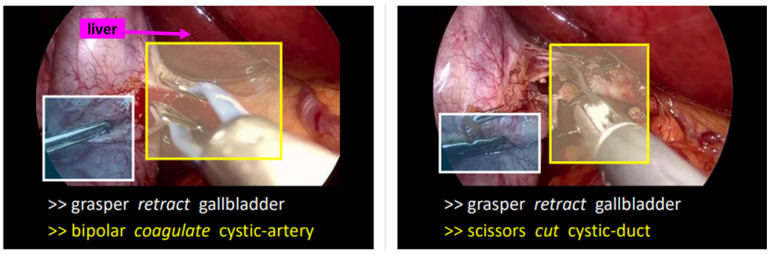

Figure 1.9: Examples of action triplets from the CholectT50 dataset. Note how, although the liver is present in both frames, it is never part of a triplet. Note also how the spatial proximity of an instrument with an anatomical structure can be used as a hint to identify the target of a triplet. Courtesy of [NYG+22].

Such considerations on *instrument-centrality* led us to define a unified framework describing the fundamental surgical CV tasks (Fig. 1.10).

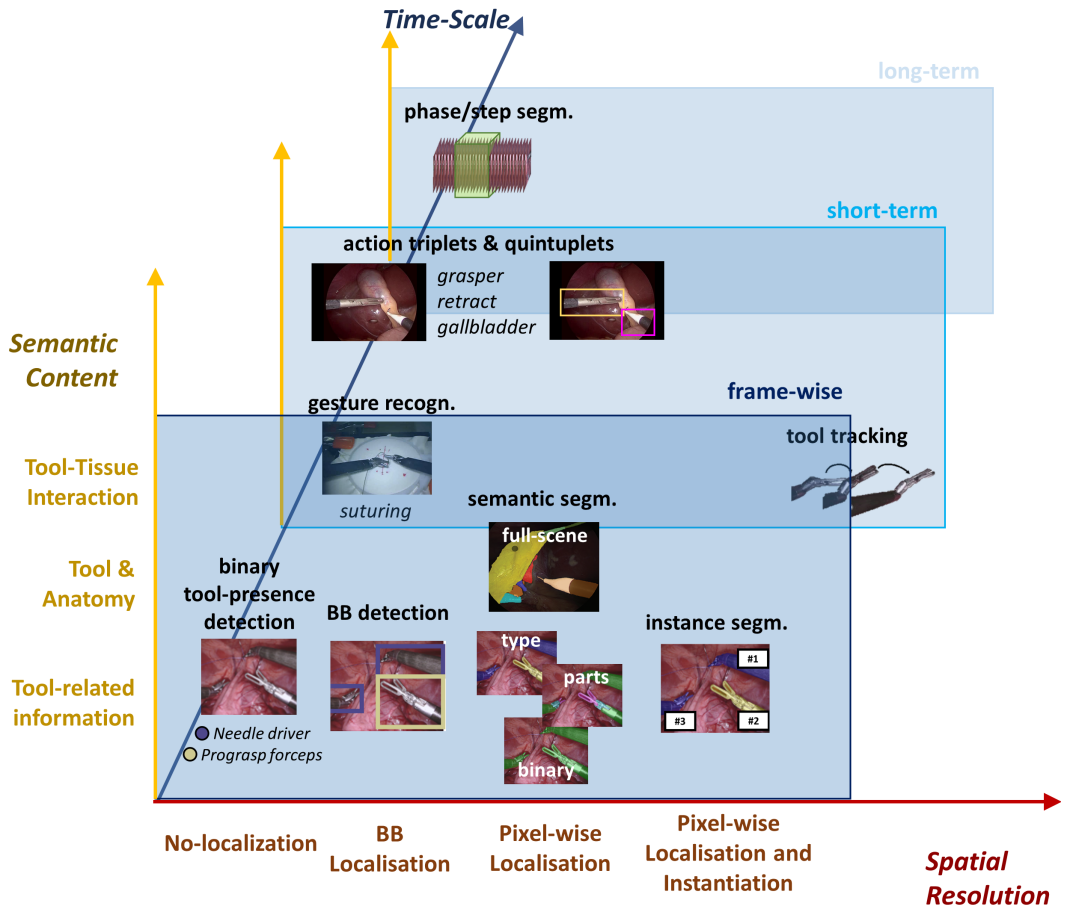

Figure 1.10: Framework collecting the main CV tasks available to digest endoscopic videos, parametrized by Time-Scale, Spatial-Resolution and Semantic Content. Frame-wise tasks include, from lowest to highest spatial resolution: tool-presence detection, bounding-box (BB) detection, semantic segmentation (binary, tool part, tool type, full-scene), instance segmentation. Short-term tasks include, from lowest to highest semantic content: action/gesture recognition, tool tracking, action triplet and quintuplet detection. Finally, long-term tasks, like step/phase segmentation.

Tasks are here parametrized by *spatial resolution*, *time-scale* and *semantic content*.

The *spatial resolution* can vary from no spatial information, as in the binary tool presence detection task, to pixel-wise localisation (e.g. segmentation).
The *time-scale* can range from single frame to multi-frames (few for action recognition, several for step/phase recognition).
The *semantic content* attribute describes the amount of semantic information extracted by the task, ranging from information describing surgical instruments only, to tool-anatomy interaction. The latter unifies different tasks like phase and triplets recognition from a semantic stand-point, considering the first a generalization of the second at a larger time-scale. For example, the surgical phase *Clipping and cutting of the cystic duct* of laparoscopic cholecystectomy [TSM+16], suggests the presence of specific interactions (*clipping*, *cutting*) between certain surgical instruments (*clipper*, *scissors*) and certain anatomical structures (*cystic duct*). This framework collects fundamental surgical CV tasks used to extrapolate dense and tractable information from endoscopic videos. This information can describe instruments exclusively (e.g. information on tool presence and localisation) or the interaction between instruments and anatomy. As previously discussed, the recognition of such interactions can be guided by instrument localisation and identification.

1.3.2 From Building-Blocks to Applications

The above-described surgical CV tasks allow to digest endoscopic videos, extracting meaningful and dense information which can enable a wide range of downstream applications. The value of instrument-related information can be also appreciated by considering such applications, including:

- **Surgical Skill Assessment**: as surgical skills are correlated with clinical outcomes, automatically assessing them represents an efficient way to provide systematic feedback and continuous training to surgeons. Although automatic skill assessment can be efficiently performed from kinematic data, this information is not available for laparoscopic surgery, the most largely adopted paradigm in minimally invasive surgery. Automatic skill assessment from videos is therefore an appealing alternative, which has been tackled by multi-stage pipelines involving tool identification, localisation and tracking as the first crucial step [LZK+21].

- **Augmented reality**: surgical augmented reality can be used as a tool to intuitively transfer additional information to the surgeon intra-operatively. In laparoscopic surgery this can be used, for example, to visualize hidden anatomical structures as overlays on the endoscopic video observed by the surgeon. A critical concern when displaying such augmented images is not occluding instruments, which can be achieved through their image-level localisation [TPPV21].

- **3D surgical scene reconstruction**: 3D reconstruction of the surgical site from endoscopic images is a prerequisite for several downstream clinical applications, including intra-operative navigation, surgical simulation, education, and robotic automation. A critical challenge of this task is the presence of instruments occluding soft tissues, which affects the completeness of surgical scene reconstruction. The problem of 3D surgical scene reconstruction has been recently tackled using Neural

Rendering. The occlusion problem was solved in [WLFD22] by introducing *mask-guided ray-casting*, a modification of the standard ray-casting algorithm bypassing rays travelling through instrument pixels during training. Although in this work tool masks were extracted manually, the real-world application of this algorithm requires a systematic way to obtain segmentation masks from endoscopic videos, which can be achieved by instrument segmentation.

- **Robotic automation**: localising surgical instruments in the 3D space finds possible applications in robotic automation and semi-automation, as for dynamic motion constraints [MISF20] and visual-servoing [ZRCM+21]. While the problem of 3D pose estimation can be tackled by means of external sensors, vision-based marker-less approaches are considered a more flexible and therefore desirable alternative. In the absence of a ground truth 3D pose of the instruments, a possible solution is to rely on the information provided by image-level localisation of the instruments [PVH09].

- **3D bowel measurement**: bowel length measurement is required by several surgical procedures, including laparoscopic Roux-en-Y gastric bypass and colon surgery. Nonetheless, the use of dedicated measuring tools like rulers is uncommon in laparoscopy, with most of surgeons relying on more rudimentary methods like using instruments of known length as reference. This approach often leads to inaccurate measurements [MGP+22]. CV offers tools to automate this process: [WMB+18], for example, propose an approach for image-based bowel measurement from stereo-camera images. Such an approach requires localisation of surgical instruments' tip to perform the measurement.

- **Critical events documentation**: documentation of intra-operative events is commonly based on operator-dictated reports, whose reliability has often been questioned [WST+13, SHW+10]. CV offers tools to automatically locate such events in surgical videos, making the post-operative reviewing process significantly more efficient than full-length video review. EndoDigest [MAU+21], for example, is a CV platform able to locate the critical phase of cystic duct division in full-length videos of laparoscopic cholecystectomy procedures. This is achieved by combining automatically extracted information about surgical phases and tool presence, for robust and reliable critical event documentation across different centers, despite a potential work-flow variability [MAL+22].

1.4 Surgical Instrument Localisation and Identification

The work presented in this thesis tackles the problems of automatic localisation and identification of surgical instruments from endoscopic videos. As discussed in Sections 1.3.1 & 1.3.2, instrument-related information can facilitate the development of solutions for other surgical CV tasks, and enable the development of downstream applications directly impacting surgical care.
Paragraph 1.4.1 formalizes the problem of automatic localisation and identification of surgical instruments, and introduces the main techniques commonly used to tackle them, later detailed in Chapter 2. Paragraph 1.4.2 highlights a common limitation of

such state-of-the-art approaches: the need for a ground truth supervision signal, usually obtained via manual annotation. Paragraph 1.4.3 introduces the potential opportunities offered by a more general knowledge about surgical instruments to solve such tasks, significantly more cost-effective compared to manual annotations. Such opportunities are extensively explored in Chapters 3, 4 & 5, with the aim to develop approaches for instrument localisation and identification free from manual annotations.

1.4.1 Problem Statement

The problem of instrument localisation is explored in this thesis both at the image level, in the form of image segmentation, and in the 3D space, in the form of vision-based 3D pose estimation. Although the two problems are intrinsically connected, and several approaches have proposed hybrid problem formulations to tackle them, they are here separately presented to better highlight individual requirements and constraints, as well as the link between them.

1.4.1.1 Image-level localisation through image segmentation

Surgical Computer Vision (CV) offers different ways to formalize the problems of image-level instrument localisation and identification, such as binary tool presence detection, bounding-box localisation and image segmentation, introduced in Figure 1.10. Among these different formalizations, image segmentation offers the possibility to simultaneously identify tools and precisely localise them in the image space.
Image segmentation is the CV task allowing to partition an image into non-overlapping segments, groups of pixels sharing common semantic attributes. Image segmentation allows to significantly simplify image representations by suppressing information outside a desired semantic set, while preserving spatial information, like object boundaries and relative object position in the image space.
Given an image $I \in R^{W \times H \times 3}$, image segmentation can be formalized in three different ways:

- **semantic segmentation:** each one of the $W \times H$ pixels of I gets assigned to a semantic label out of a predefined set $\{S_i\}$, with i in $[0, N_{cls}]$ - where 0 is the *background* class - which defines the semantic scope of the task:

$$SeS : I \in R^{W \times H \times 3} \rightarrow I_{SeS} \in R^{W \times H \times N_{cls}+1}. \tag{1.1}$$

  Table 1.1 reports some popular applications of semantic segmentation in the surgical CV domain, highlighting their semantic scope, as well as popular datasets annotated for them. Note that the presented variants are normally referred to as separate classes of problems: *binary segmentation*, *tool part segmentation*, *tool type segmentation*, *anatomy segmentation* are all highly researched problems, having their specific datasets and dedicated approaches. Examples of such tasks are shown in Figure 1.11, columns 2-4.

  Semantic segmentation formalization has been highly adopted in recent years, concurrently with the rise of deep learning, as this formulation allows to solve the problem with minimal modifications to standard neural network architectures

	Task	Scope	Datasets
Semantic	Binary	Separate *tool* pixels from *anatomy* pixels	EndoVis 2017,2019
	Tool Part	Separate pixels based on the *tool* part they belong to	EndoVis 2015,2017
	Tool Type	Separate pixels based on the type of *tool* they belong to	EndoVis 2017
	Anatomy	Separate pixels based on the *anatomy* they belong to	HeiSurf
	Full-Scene	Separate pixels based on the type of *tool* or *anatomy* they belong to	EndoVis 2018, CholecSeg8k, CaDIS, Endoscapes
Instance	Tool Part	Separate instances based on the *tool* part they represent	
	Tool Type	Separate instances based on the *tool* type they represent	
Pan.	Full-Scene	Separate instances or pixels based on the type of *tool* or *anatomy* they represent/belong to	

Table 1.1: The table reports an overview of the main surgical instrument segmentation tasks, organized according to their formalization in Semantic, Instance and Pan-Optic (Pan.). Each task is characterized by its semantic scope. The main datasets annotated for each task are also reported. Datasets included are the segmentation datasets of the MICCAI EndoVis challenges 2015, 2017 [ASK+ 19], 2018 [AKB+ 20], 2019 [RRF+ 20], CaDIS [GFK+ 19], HeiSurf; CholecSeg8k [HKK+ 20]; Endoscapes [AMV+ 21].

developed and tested for image classification. Indeed, a significant segment of literature for surgical instrument segmentation is based on encoder-decoder architectures directly learning the mapping SeS between images and segmentation masks from datasets manually annotated with pixel-wise semantic labels [GPHLF+17, SRKI18a, PPA+19, HL19]. However, more recent works have started questioning the suitability of pixel-wise classification for tool type segmentation. Tool parts like shafts, in fact, are often similar across different instrument types, making pixel-wise classification challenging to solve. In addition, semantic segmentation is not suitable to distinguish between separate tool instances sharing the same semantic label. These limitations are addressed by instance segmentation;

- **instance segmentation:** the image I is partitioned into a certain number N_{Inst} of tool instances, not a-priori specified. Each instance can be represented by a binary mask $M_i^{Inst} \in R^{W \times H}$, which is then assigned as a whole to a label out of the predefined set $\{S_i\}$, with i in $[1, N_{cls}]$:

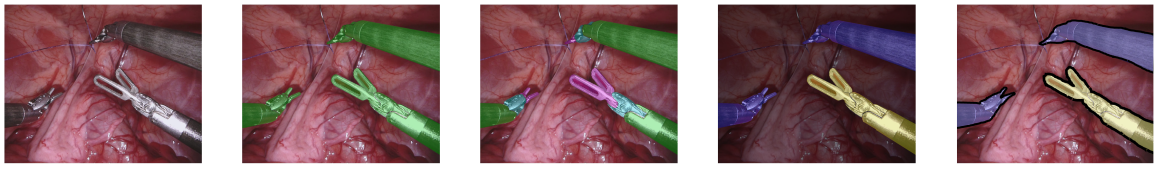

Figure 1.11: From left to right: original endoscopic frame, binary segmentation, tool part semantic segmentation, tool type semantic segmentation and tool type instance segmentation (separate instances highlighted by their boundary).

$$IS : I \in R^{W \times H \times 3} \rightarrow \{M_i^{Inst} \in R^{W \times H}\}, \{S_i \in [1, N_{cls}]\}, \text{ with } i \text{ in } [1, N_{Inst}]. \tag{1.2}$$

As for semantic segmentation, the set $\{S_i\}$ defines the semantic scope of the task. Instance segmentation, for example, could be applied to both tool parts and tool types. However, following the common use of the term in the surgical CV community, we will refer to tool type instance segmentation simply as *tool instance segmentation*.
Instance segmentation (Figure 1.11, last column) shifts the focus from pixel-wise classification to the concept of *instances*, as commonly done for object detection problems. Indeed one of the most popular solutions to instance segmentation is Mask-RCNN [HGDG17], a straightforward extension of RCNN model for object detection [GDDM14]. This solution has been widely adopted by state-of-the-art approaches [KJD+21, GBSA20] trained on data manually annotated with pixel-wise semantic and instance labels. Instance segmentation formalization calls for the definition of the concept of *things*, the countable objects, and *stuff*, uncountable objects. While the distinction between the two is to some extent subjective, some objects (e.g. *sky,grass* in natural images, or *fatty tissue* in endoscopic images) are clearly uncountable, and fall outside the scope of instance segmentation. This limitation is addressed by pan-optic Segmentation;

- **pan-optic segmentation:** the image is simultaneously treated under both the instance and semantic segmentation lenses. *Things* are instantiated and classified as in instance segmentation, while *stuff* undergoes direct pixel-wise classification. This approach provides the most complete segmentation formalization, and is currently highly researched in the general CV community [LC22]. To the best of our knowledge, no approach has yet been proposed for surgical CV, and no dataset has been explicitly annotated for this task, so its use is not further discussed in this thesis.

Since surgical instruments are countable objects - *things* - both semantic and instance segmentation are suitable formulations. Note that an instance segmentation representation can be transformed into a semantic segmentation representation, but not vice-versa, as the instantiation information would be missing from the semantic mask:

$$\{M_i^{Inst} \in R^{W \times H}\}, \{S_i \in [0, N_{cls} - 1]\}, \text{ with } i \text{ in } [1, N_{Inst}] \quad \rightleftarrows \quad I_{SeS} \in R^{W \times H \times N_{cls}+1}. \tag{1.3}$$

A complete overview of existing segmentation approaches is presented in Chapter 2.

1.4.1.2 Localisation in the 3D space

Beyond localisation at the image level, knowing the 3D pose and shape of surgical instruments is often required by robotic applications, as discussed in Section 1.3.2. Such a problem can be tackled using external sensors like electromagnetic trackers and fiber Bragg gratings, or applying optical markers to the instruments [SLQ+16, CGD07, GR18, WS20]. However, relying on external sensors involves several undesirable actions to be taken, such as modifying the instrument design to fit sensors/markers, potentially requiring their re-certification. For this reason, vision-based marker-less solutions represent an appealing alternative [BASJ17].
A vision-based approach for 3D tool localisation commonly features the following elements:

- a 3D virtual/CAD model of the robotic instruments r, parametrized by n values and characterized by a set of geometric parameters. The n parameters can be for example kinematic joint values or pose parameters. The two are interchangeable and have both been explored in literature. For the sake of simplicity, we will refer to the case of kinematic modelling to define the general framework;

- a set of kinematic joint values vector $\{\mathbf{k}_i\}$ with i in $[1, T]$, with T being the number of time-stamps, and each $\mathbf{k}_i$ having n components. The kinematic joint values, fed into r, give the estimated 3D shape of the robotic instruments in the instrument reference frame. In case of pose parametrization each vector $\mathbf{k}_i$ would be replaced by a pose vector;

- a camera - robot base transformation b, describing the relative pose between the endoscopic camera reference frame and the robotic instrument base reference frame;

- a camera model c, characterized by a set of extrinsic and intrinsic parameters. If known, the camera is said to be *calibrated*. For simplicity, we will use c to describe the projective transformation mapping a set of 3D points describing the instrument 3D shape into the rendered tool image $\hat{m}$;

- a set of images $\{I_i\}$, with i in $[1, T]$, paired with the set of kinematic values $\{\mathbf{k}_i\}$;

- a couple of transformations P_{real}, P_{rend}, mapping the real image I_i and the rendered tool image $\hat{m}_i$ into the same projection space.

The combination of the listed elements allows to map a kinematic joint values vector $\mathbf{k}_i$ into the corresponding rendered tool image $\hat{m}_i$:

$$\hat{m}_i = c(br(\mathbf{k}_i)). \tag{1.4}$$

Usually, kinematic values recorded by robotic systems tend to be inaccurate, due to two main reasons: 1) tool-tissue interaction can modify the configuration of the instruments,

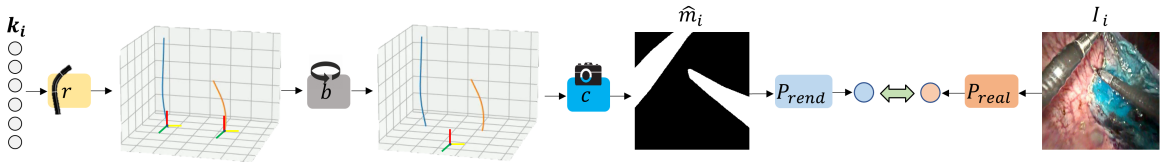

Figure 1.12: Overview of the general framework commonly used to tackle a vision-based 3D tool localisation problem.

with respect to the one specified by the user and recorded by the robotic system; 2) unmodelled non-linearities in the tool model (e.g. cable friction, slack in instrument channel, backlash) can lead to loss of motion between the motors and the instruments: if not properly modelled this can result in a significant mismatch between estimated tool configuration and effective one. These two factors lead to a non-deterministic mapping between recorded kinematics and actual robot configuration, which makes recorded kinematics unreliable. For this reason, as shown in Figure 1.12, a vision-based 3D pose estimation problem is often formulated as an optimization problem aimed at minimizing the distance between the projections into a common space of the real image, through P_{real}, and of the rendered tool image, through P_{rend}:

$$P_{real}(I_i) = P_{rend}(\hat{m}_i). \tag{1.5}$$

Conceptually, this hybrid framework ties together vision-based instrument localisation and 3D pose estimation, and has been used to tackle both problems [AOH+18, dCRPR19]. Different approaches which have adopted this hybrid framework [AOH+18, DZK+22] formalize P_{real} as machine/deep learning model, whose parameters are learnt from labelled data, commonly annotated for the tool segmentation task. The most relevant solutions adopting such frameworks are described in Chapter 2.

1.4.2 The Annotation Bottleneck

As introduced in the previous section, Deep Learning (DL) is the current method of choice to tackle the problem of instrument localisation and identification. More generally, over the last decade, DL, fueled by a constantly increasing amount of data and computing capabilities, has outperformed standard CV algorithms in a variety of tasks like object detection, image segmentation and classification [VDDP18]. Compared to standard CV algorithms, which require hand-crafted feature selection and extraction, DL lets a neural network function learn its own parameters from a set of training data, without being explicitly programmed. In the largely used *fully-supervised* paradigm, this is obtained by optimizing neural network parameters with the greedy objective of yielding predictions matching a manually annotated ground truth. While this paradigm does not come without shortcomings, such as the possible over-fitting of the training data, it quickly led to breakthrough results in CV. Specifically, DL algorithms outperformed standard CV algorithms in all those tasks requiring a semantic understanding of data, where factors like large *intra-class* variability, or reduced *inter-class* variability, make the process of manual feature selection non-obvious or, at least, extremely inefficient.

In the surgical CV field, the mental process followed by experts to carry out tasks

is developed through years of experience, and is often hard to explicitly formalize: explicitly defining relevant features for standard CV algorithms development is therefore prohibitive. Motivated by this challenge, and fueled by the excitement of initial breakthrough results [TSM+16], the fully-supervised approach has been widely used for most surgical CV tasks, including surgical instrument localisation and identification. However, this *one-fits-all* approach, requiring manual annotations to solve every surgical CV problem, appears today unsustainable and incompatible with clinical translation. In order to guarantee robustness and generalization ability, in fact, DL approaches need to be trained on large amounts of data, capturing the potential variability of real-world data. However, as the size of datasets increases, the cost of annotation linearly increases with it. Building Endoscape [AMV+21] for example, a dataset for full-scene segmentation of endoscopic images, required over 400 hours of work for 2k images. Annotations were performed in double by multiple computer scientists and surgeons, as problem-specific knowledge was required from both sides. The need for such domain expertise significantly increases the cost of annotations.
Regardless of the individual laboratories' or companies' resources, such cost of annotation can quickly become unaffordable. This creates an *annotation bottleneck*, confining DL model training to a tiny fraction of the potentially available data. If not properly addressed, this *annotation bottleneck* can severely limit the benefits that surgical big data could bring to surgery, by negatively impacting research and translation in different ways:

- **lack of model generalization ability**: the most immediate effect of training in small datasets is the inability of models to perform well on unseen data, due to unwanted biasing factors present in the collected data. This can pose severe problems for translation of the developed technologies. A clear example of this problem in the healthcare sector is the infamous failure of AI to provide reliable predictive tools during the Covid-19 pandemic [Cha22];

- **limited benchmark dataset representativeness**: the highly-competitive nature of research [Žel23] pushes researchers to develop solutions aimed at outperforming state-of-the-art approaches on specific benchmark datasets. Such datasets allow standardized evaluation of methods, promoting fair comparison among them. However, when the number and the size of such benchmark datasets is reduced because of the *annotation bottleneck*, their ability to represent real problems may be hindered. This can lead to the development of methods over-optimized to perform well on potentially non-representative datasets, thus failing to effectively advance the state-of-the-art;

- **tasks compartmentalization**: the lack of centralized data collection and annotation in the surgical data science community [MHES+22], has commonly led to the construction of several task-specific datasets, tailored to individual laboratories' needs. Although some exceptions exist, like CholecSeg8k and CholecT50 datasets [HKK+20, NYG+22], re-annotating existing datasets for different tasks is uncommon. This can have the long-term effect of compartmentalizing surgical CV tasks, as researchers are forced to find solutions to problems without relying on information coming from different tasks. This is extremely counter-intuitive for surgical CV tasks, which, as described in Section 1.3.1, often tackle the same problem (e.g.

surgical work-flow analysis) from different perspectives (e.g. gesture recognition, action-triplet identification, phase/step segmentation), all potentially benefiting from the availability of the same information (e.g. tool-related information).

In order to mitigate the *annotation bottleneck* problem, different solutions have been proposed in the general CV community. *Semi-supervised* learning approaches, for example, incorporate unlabelled data in the training process, while still requiring access to a set of manually annotated data. Similarly, *domain adaptation* techniques allow to repurpose knowledge learnt from a certain domain to another domain, featuring few or no annotated samples. Recently, *self-supervised* representation learning approaches have gained significant traction in the community. These approaches aim at learning useful feature representations of data, capturing their semantic content, with no external supervision. Such representations can then be used to learn downstream tasks from small sets of annotated data, allowing to boost models performance, reduce training time and minimize the amount of manual annotations needed.
Such strategies have been successfully applied to several surgical CV tasks, including instrument localisation and identification [RSA+22, AMV+21, SJDH21, ZJG+20]. However, we believe that the surgical domain offers specific opportunities to solve the tool localisation and identification tasks in a completely unsupervised way, which have been only marginally explored in literature.

1.4.3 The Potential of Prior and Complementary Knowledge

This thesis explores the use of prior and complementary knowledge about surgical tools to perform unsupervised instrument localisation and identification from endoscopic videos. Such knowledge can come in different forms, and derives from observations about surgical tools and the surgical domain in general, reported below:

- ***standardization* of surgical instruments**: surgical instruments' shape and overall appearance is almost completely determined by their function. Therefore, the appearance variability arising, for example, from different manufacturers, is usually minimal, with key instrument features always well-preserved. This contrasts with the large intra-class variability found in general CV problems (Fig. 1.13). Such standardization allows to build prior models of the instruments. In this thesis, we distinguish between *weak*, more general prior models, like template images of the instruments (*shape-priors*, Figure 1.14, left) and *strong* prior models, like parametrized 3D virtual/CAD models (Figure 1.14, right);

- **constraints on instrument motion**: surgical instruments' motion inside the surgical site is constrained and characterized by two main factors. First of all by physical characteristics of the instruments: laparoscopic instruments are usually rigid, therefore their motion is coherent as opposed to the one of surrounding soft tissues (Fig. 1.15, left). Instruments' motion is also characterized by basic principles imposed by the surgical technique or implicitly followed by the surgeons handling them. In laparoscopy, for example, the principle of *triangulation* tends to make instruments enter the field of view from the lateral side; in addition, surgeons tend to avoid surgical instruments overlap, in order to reduce the chances of mutual tool

Figure 1.13: Top row shows randomly sampled images from the ImageNet [DDS+ 09] chair class. Bottom row shows images of different types of laparoscopic scissors, produced by different manufacturers. The scissors, whose appearance is totally constrained by their function, tend to keep clearly recognizable features regardless of the specific type and manufacturer.

Figure 1.14: Left: shape-priors, in the form of binary segmentation masks of the instruments, obtained from automatic segmentation of chroma-key images. Courtesy of [GPHFD+ 21]. Right: different views of a 3D CAD model of a laparoscopic forceps head. Courtesy of [Gra]

occlusions and unwanted tool interactions;

- **availability of multi-modal complementary information**: endoscopic videos are not the only source of online information capturing surgical instruments' activity during procedures. For example, electric signals from active instruments, like bipolar forceps, can be used to detect their presence in the surgical scene. In robot-assisted surgery, instrument usage can be automatically recorded by robotic systems, as well as instrument motion through robot kinematics. Furthermore, manually annotated information about the surgical workflow, like phases, steps and binary tool presence, can be considered as a source of complementary knowledge describing surgical tool activity. We define these different sources of multi-modal information as *complementary* with respect to the tasks of tool segmentation and 3D pose estimation, as they do not directly provide ground truth information to solve them;

- **simulation**: simulators offer the possibility to young surgeons to build experience by training in a safe environment. Procedure-specific virtual reality simulators

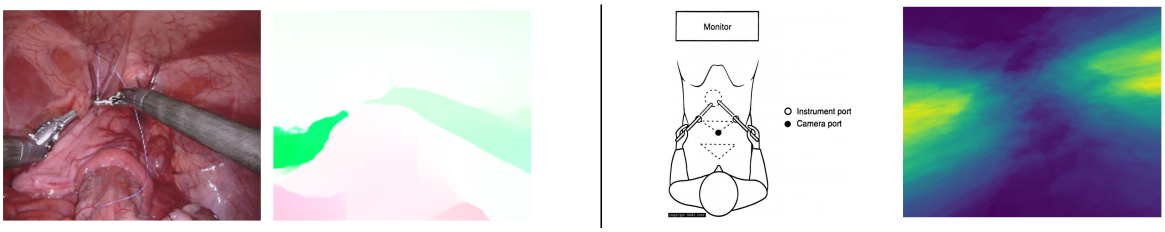

Figure 1.15: Left: graphic visualization of instrument coherent motion vs soft tissue incoherent motion using optical flow. Right: laparoscopic triangulation principle for trocar placement (courtesy of [SAG14]), and heat-map representing instrument localisation in the image space during a procedure extracted from the MICCAI 2017 grand-challenge dataset on instrument segmentation [ASK+ 19] (blue low presence, yellow high presence). Note how the two instruments do not normally overlap and tend to enter the field-of-view from the sides.

like LapSim [DHM+05] can provide accurate renderings of different surgical sites and instruments, as well as the interactions between the two (Fig. 1.16). Being synthesized, the surgical scene in simulators is known in every aspect and can be used to generate realistic labelled samples for DL algorithms training.

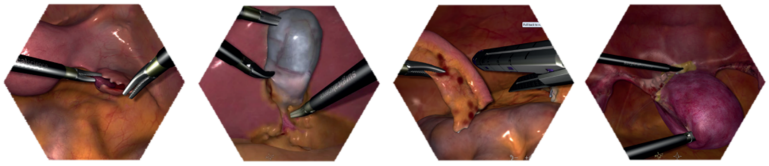

Figure 1.16: Renderings from the LapSim® simulator: from left to right salpingectomy, cholecystectomy, appendectomy and hysterectomy procedures. Courtesy of [Lap20].

1.4.4 Research Question

In contrast to standard manual annotations used to tackle the tool localisation and identification tasks (e.g. bounding-boxes, segmentation masks), prior and complementary knowledge provide cheaper and more flexible information about the problem. Prior knowledge, in particular, is usually not directly linked to a specific set of data. Information about tool shape and color distribution, for example, can be applied across surgical domains, for data collected from different procedures, performed with different techniques and in different centers. Complementary knowledge, while usually linked to a specific set of data, can often be obtained automatically (e.g. kinematics and tool usage from robotic systems), or with less annotation effort (e.g. binary tool presence and phase/step labels) compared to standard manual annotations.
These features make prior and complementary knowledge a more general and repurposable source of information, which can allow models to learn from unlabelled data, without incurring in the *annotation bottleneck* problem. On the downside, integrating such information in standard DL architectures is not straightforward, compared to standard fully-supervised training (Figure 1.17).
The hypothetical framework shown in Figure 1.17 (right) poses interesting research questions: *how can general knowledge about surgical tools be formalized into a pseudo-*

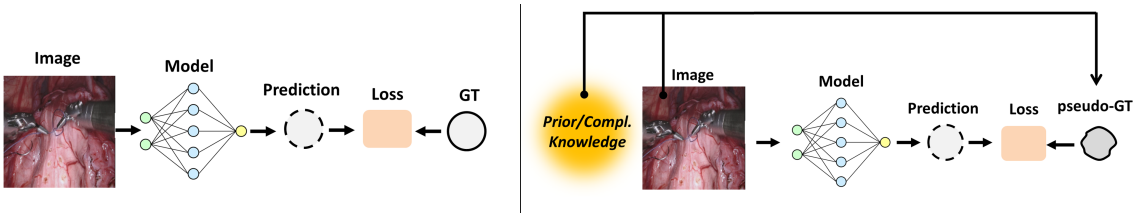

Figure 1.17: Left: fully-supervised learning framework, where a ground truth signal (GT), usually obtained through manual annotation, is directly used to compute the loss for model training. The model's parameters are optimized to minimize such loss, so that its predictions can match the GT. Right: hypothetical framework using prior/complementary knowledge to generate a pseudo-GT signal for deep learning model training. This framework opens up interesting questions, such as how to generate the pseudo-GT from prior/complementary knowledge and how to effectively learn from it.

supervision signal? How can a DL model effectively learn from such a signal?
Our contributions, presented in Chapters 3, 4 & 5, address these questions, exploring the use of such knowledge to tackle the problems of instrument localisation and identification from endoscopic videos.

1.5 Thesis Contribution

In this thesis we investigate the problems of surgical instrument localisation and identification, both at the image level, in the form of segmentation, and in the 3D space, in the form of 3D pose estimation. Motivated by the need to overcome the *annotation bottleneck* problem, discussed in Section 1.4.2, prior and complementary knowledge about tools is incorporated in the developed architectures, to replace manual annotations. This yields annotation-free approaches, trainable on unlabelled data. Our three main contributions (Fig. 1.18) are introduced below and detailed in Chapters 3, 4 & 5.

Contribution 1, FUN-SIS: a Fully UNsupervised approach for Surgical Instrument Segmentation:
As our first contribution, we design an approach for unsupervised binary instrument segmentation. The proposed solution trains on completely unlabelled videos, exploiting prior knowledge of instrument appearance and motion. Prior knowledge on tool appearance is formalized as shape-prior masks, binary segmentation masks of surgical instruments, obtainable in various ways such as *recycling* existing annotations from different datasets or projecting 3D tool models in the image space. For instrument motion, a simple assumption is made: compared to the surrounding anatomy, surgical instruments move coherently, i.e. two points close-by in an instrument normally move in the same direction. This knowledge is incorporated into a deep learning architecture and used to produce a pseudo-supervision signal to supervise binary segmentation training. The signal is further refined by exploiting peculiar properties of neural networks when dealing with noisy labels. We validate this approach in different datasets, using different kinds of shape-priors, achieving results comparable with fully-supervised solutions.

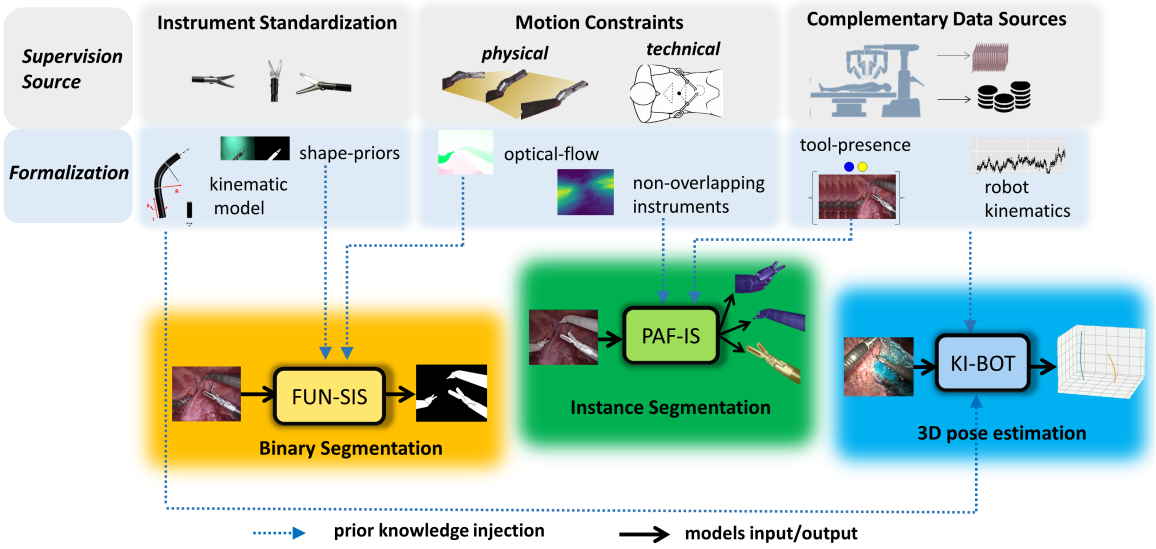

Figure 1.18: Overview of the contributions proposed in this thesis, highlighting the sources of information exploited to replace manual annotations for deep learning model training. Our first contribution (FUN-SIS) tackles the binary segmentation problem, using instrument shape-priors and the hypothesis of instrument coherent motion as sources of prior knowledge. Our second contribution (PAF-IS) builds on top of FUN-SIS to solve the instance segmentation task by exclusively relying on general hypotheses on instrument positioning in the field-of-view and binary tool presence labels. Finally, our third contribution (KI-BOT) uses instrument kinematic modelling to learn 3D pose estimation from inaccurate kinematic data only.

Contribution 2, PAF-IS: a Pixel-wise Annotation Free framework for Instance Segmentation of surgical tools:
As our second contribution, we propose a novel framework for instance segmentation model training, designed to minimize human annotation effort by removing the need for pixel-wise semantic and instance annotations. Without an explicit supervision signal, our solution learns to extract individual tool instances from binary segmentation masks, and obtains, for each tool instance, a powerful feature representation via self-supervised contrastive learning. Such instance-wise representations guide the automatic selection of a tiny number of instances (as few as 8 in our experiments), displayed to a potential human user for tool type labeling. The gathered information, in combination with binary tool presence labels, guides the training of an instance-wise classifier, predicting a tool type label for each tool instance.

Contribution 3, KI-BOT: a Kinematic Bottleneck Approach For Pose Regression of Flexible Surgical Instruments directly from Images:
As our last contribution, we leverage the availability of strong prior knowledge about instruments by proposing the introduction of parametrized 3D tool models as part of end-to-end trainable DL pipelines for direct 3D pose estimation from endoscopic images. The proposed framework trains a model to directly predict kinematic joint values from images, by exclusively relying on automatically recorded kinematic data. We design our approach to be robust to noisy kinematic data, and validate it in the challenging domain of flexible endoscopic surgery.

1.6 Outline

The work presented in this thesis is organized according to the following outline:

- Chapter 2 provides a structured overview of previous publications relevant to the tasks of surgical instrument segmentation and 3D pose estimation, organizing them according to their learning methodology.
- Chapter 3 presents FUN-SIS, an approach for binary instrument segmentation, training on unlabelled data exploiting prior knowledge on instrument shape and motion.
- Chapter 4 presents PAF-IS, a novel framework allowing to train an instance segmentation model, exclusively relying on prior knowledge of instrument positioning in the image space and weak complementary information in the form of binary tool presence labels.
- Chapter 5 presents KI-BOT, a framework for 3D pose estimation integrating a parametrized 3D model of surgical instruments inside an end-to-end trainable deep learning architecture for direct kinematics regression from images.
- Chapter 6 first summarizes the work presented in this thesis, presenting a unified framework to learn the tool localisation and identification tasks from unlabelled datasets. This framework helps us discuss the analogies between the proposed contributions and contextualize them with respect to relevant deep learning paradigms, like self-supervised representation learning and learning-from-noisy-labels. Finally, current limitations of the proposed solutions, paths for further development and open questions arising from our work are discussed.

CHAPTER 2

Related Work on Surgical Tool Localisation and Identification

Contents

2.1 Fully-Supervised Solutions 48

2.2 Semi-Supervised Solutions 48

2.3 Prior and Complementary Knowledge based Solutions 49

2.3.1 Simulation and Semi-synthetic Data Generation 50

2.3.2 Complementary Information based Solutions 51

2.3.3 Weak Prior Knowledge based Solutions 52

2.3.4 Strong Prior Knowledge based Solutions 54

2.4 Thesis Positioning 56

This chapter presents related literature on instrument localisation and identification, both at the image level, in the form of image segmentation, and in the 3D space, in the form of vision-based 3D pose estimation.
Existing solutions are organized according to their core learning methodology: fully-supervised solutions, completely relying on manual annotations, are presented in Section 2.1; semi-supervised solutions, combining labelled and unlabelled data in the training process, are discussed in Section 2.2. Prior and complementary knowledge based solutions are finally presented in Section 2.3.

2.1 Fully-Supervised Solutions

Following the Deep Learning (DL) breakthrough in the field of surgical computer vision, marked by seminal works like Endo-Net for workflow analysis [TSM+16], research works have mostly addressed the problem of surgical tool segmentation using fully-supervised DL approaches. Such approaches have largely outperformed previously proposed methods for the tasks of instrument segmentation and tracking [BAA+18]. In particular, encoder-decoder architectures based on Convolutional Neural Networks (CNNs) have been widely adopted, in concurrency with a semantic segmentation formulation of the problem. [GPHLF+17, SRKI18a, PPA+19, HL19] propose different variations of U-Net architecture [RFB15], exploring different loss functions, residual connections, dilated convolutions and ad-hoc augmentation pipelines. Multi-task learning has also been adopted, coupling the segmentation task with image-based localisation of tool landmarks [LRR+17] and task-oriented saliency maps prediction [IVLR21]. While the segmentation task can be solved from single frames, temporal information has been proven to boost performance, especially in the case of partially occluded tools [JCDH19a].

Recently, instance segmentation approaches have started gaining traction. As previously discussed, the instance segmentation formulation is particularly suitable to deal with surgical instruments, and allows to extract reacher information compared to semantic segmentation approaches. The proposed approaches to tackle this problem require pixel-wise semantic and instance labels to train. Most of them are based on the popular Mask-RCNN architecture [HGDG17]. [KJD+21] directly train a Mask-RCNN architecture for the task of surgical instrument instance segmentation, and validate it across different datasets, highlighting the importance of cross-dataset training to improve robustness and generalization ability (Figure 2.1). ISI-Net [GBSA20] adds a temporal-consistency module for improved segmentation results. Beyond Mask-RCNN based approaches, [KMNA+21] use an anchor-free approach for instrument instantiation, based on direct prediction of instruments centroids position. [ZJH22] simultaneously tackle the problems of instance segmentation and tracking, using a transformer-based architecture.

2.2 Semi-Supervised Solutions

This family of approaches incorporates unlabelled data in the training process, while still requiring access to a set of manually annotated data to supervise the training. Different

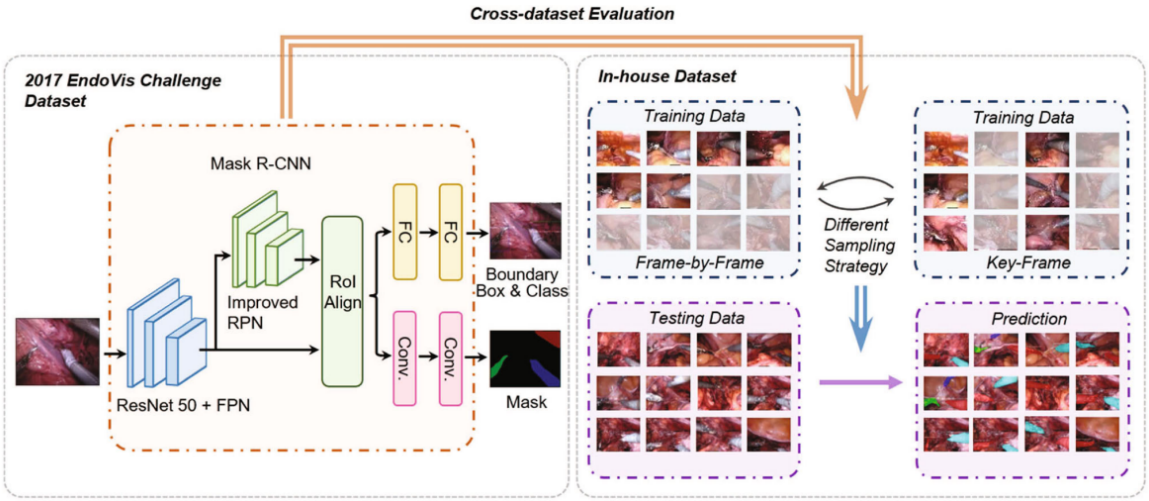

Figure 2.1: Overview of [KJD+21] approach and validation. The approach consists of a standard Mask-RCNN architecture, validated using cross-dataset evaluation with different sampling strategies. Courtesy of [KJD+21].

solutions to combine unlabelled and labelled data have been explored. [RZV+18] pre-train a segmentation model on unlabelled data, by means of a re-colorization pretext task carried out using a cycle-GAN architecture, and then fine-tunes the model on annotated data (Figure 2.2). A similar pipeline can be followed by replacing the pre-text task with self-supervised representation learning on the unlabelled data, as experimentally shown by [RSA+22]. In their work, different state-of-the-art self-supervised representation learning approaches, both contrastive-based, like MOCO [HFW+20] and distillation based, like DINO [CTM+21], were bench-marked on different surgical computer vision tasks, including instrument segmentation. Their analysis shows that substantial performance gains can be achieved using self-supervised pre-training over common ImageNet initialization. [ZJG+20] tackle the problem of sparsely annotated data, propagating low hertz annotations to intermediate unlabelled frames using optical flow. [KAN+21] incorporate in the training process unlabelled data from different domains, to improve generalization to those domains. This is achieved by mapping annotated frames from the labelled set to the unlabelled domain using a cycle-GAN architecture, allowing for better generalization.

2.3 Prior and Complementary Knowledge based Solutions

As discussed in Section 1.4.3 prior problem knowledge and complementary multi-modal information can be used to directly tackle the problem of instrument localisation and identification, minimizing the need for manual annotations. The solutions here presented are organized based on the main source of information they use to solve such problems. Solutions exploiting simulation and data synthesis are described in Section 2.3.1. Solutions using complementary information about surgical instruments, like binary tool presence, are presented in Section 2.3.2. Solutions using weak prior knowledge, in the form of shape-priors or general assumptions on instrument motion and color appearance, are described in Section 2.3.3. Finally, solutions using strong prior knowledge,

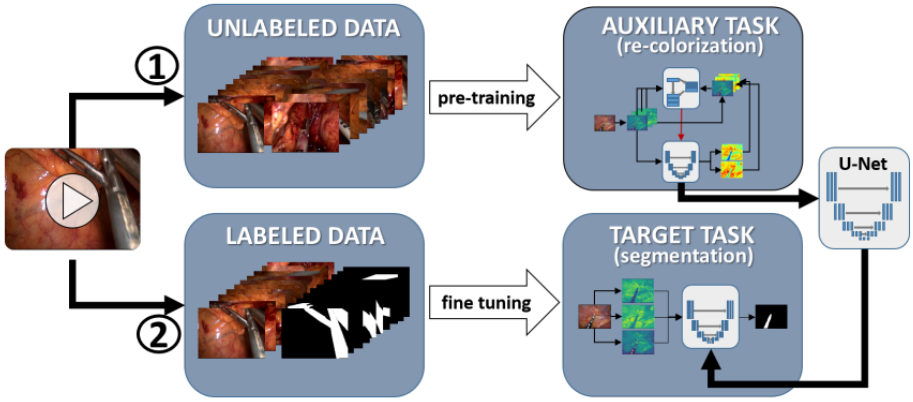

Figure 2.2: Overview of [RZV+ 18] approach. The approach includes a pre-training step, performed using a pretext re-colorization task on unlabelled data, and fully-supervised training step, on the available labelled data. Courtesy of [RZV+ 18].

in the form of parametrized 3D models of the instruments, are discussed in Section 2.3.4. It is worth pointing out that several of the presented works combine different sources of information, as this helps achieve better and more robust results. Table 2.1, at the end of the Chapter, collects all the methods discussed in this Section, highlighting the different sources of information used by each of them.

2.3.1 Simulation and Semi-synthetic Data Generation

Full surgical scene simulators allow to generate virtually infinite-sized datasets of synthetic endoscopic videos with known ground truth information about anatomy and surgical tools. However, directly learning from synthetic images is prohibitive, as there still exists a significant domain gap between real and simulated data. In order to bridge this gap different solutions have been proposed. [PFR+19] exploit a cycle Generative Adversarial Network (cycle-GAN, [ZPIE17]) to translate simulated images into real-looking laparoscopic images, while preserving their original content (Figure 2.3). Fully-supervised learning is then carried out on the generated semi-synthetic data. [SSMZ20] propose Endo-Sim2Real, a consistency-based framework for joint training from simulated and unlabelled real data. This approach does not explicitly perform image-to-image translation, but trains a model for instrument segmentation using a fully-supervised loss on simulated images, and a consistency loss on different augmented versions of real images. [SMZ21] improves such framework employing a teacher–student paradigm, developed to address the confirmation bias problem affecting the consistency loss of Endo-Sim2Real.
However, full-scene simulators are not always available. The CoppeliaSim DaVinci simulator [FMFV22] for example, allows to perform non-surgical tasks, like object manipulation. Therefore, renderings of the surgical scene cannot be generated. Simulated images of surgical instruments are used in [CS21], which map them to realistic-looking ones using a Cycle-GAN. Manually annotated segmentation masks are needed to produce the real domain set for Cycle-GAN training. Translated tool images can then be pasted on background-only images to generate semi-synthetic samples with known ground truth tool masks.

In the absence of simulators, semi-synthetic samples have been produced by [GPHFD+21], by merging automatically segmented tools from green-screen recordings and real surgical background images (Fig. 2.4). Finally, [MMC+21] use a *pix2pix* GAN to generate synthetic endoscopic images from the ground truth segmentation masks of surgical instruments and anatomy.
As the quality of simulators improves, simulation-based approaches will play a crucial role to alleviate the burden of manual annotation. However, as for now, existing approaches still struggle to bridge the gap between real and simulated data, and the cost of advanced simulators still poses a significant barrier to their use.

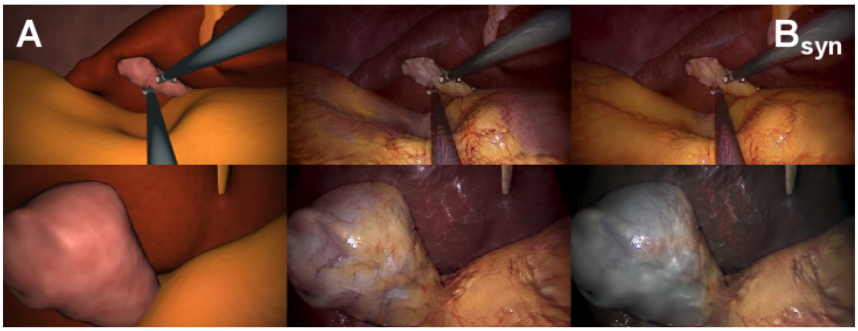

Figure 2.3: Images from full-scene laparoscopic simulation (first column) translated into real-looking laparoscopic images (second and third column) using different styles. Courtesy of [ZPIE17].

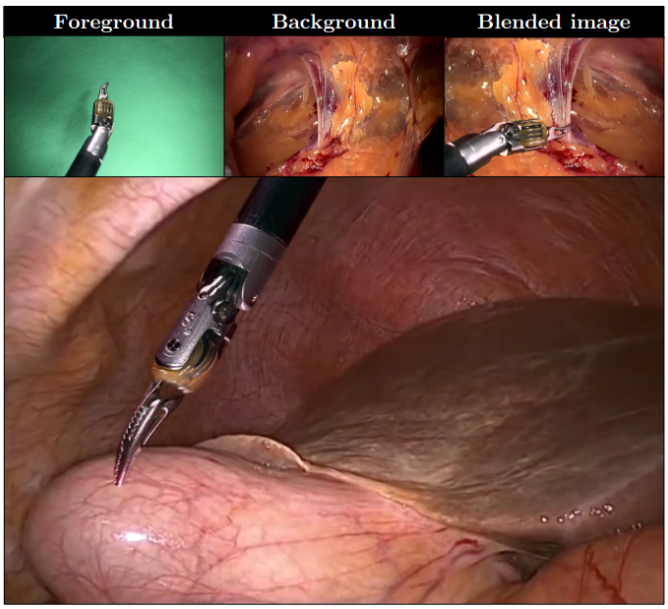

Figure 2.4: Blending process (top row) and blended image sample (bottom picture) created using green-screen recordings of surgical instruments and background-only images. Courtesy of [GPHFD+21].

2.3.2 Complementary Information based Solutions

As discussed in Section 1.4.3, complementary multi-modal information about surgical tools can be acquired in different forms. In robotic-assisted surgery, kinematic joint

values can often be recorded, as well as the type of instruments inserted in the robotic arms. Furthermore, manually annotated information about the surgical workflow, like phases, steps and binary tool presence, can be considered as a source of weak knowledge for the tool localisation task, obtainable at a significantly lower cost compared to pixel-wise annotations. As the use of kinematic joint values usually requires the availability of kinematic models of the instruments, related approaches are separately discussed in Section 2.3.4, as part of the methods using parametrized 3D tool models.
Weak annotations, in the form of binary tool presence labels describing tool presence, have been mostly used to tackle the problem of instrument detection using bounding-boxes. [VMMP18] train a multi-label classifier to predict tool presence from single frames; the designed architecture features an extended spatial pooling layer yielding class-specific feature maps, used during inference to localise the tools. Similarly [NMMP19] use Wildcat Pooling to obtain localisation maps, adding a convolution-LSTM module for improved temporal consistency. Differently from these two approaches, [XLL+22] use binary tool presence annotations, in combination with green-screen recorded images of surgical instruments, to obtain a pseudo-supervision signal consisting of noisy and redundant bounding boxes. A bounding-box regressor is then trained on the noisy supervision signal, and its predictions for a certain tool are averaged together according to their confidence score. For the task of segmentation, [YSL22] automatically obtain a pseudo-supervision signal by attaching an electromagnetic sensor to surgical instruments. While cutting the cost of annotations, the approach is inherently limited by regulatory constraints, which limited the extent of validation of that study.
The use of binary tool presence annotations has remained limited to the localisation task, as the standard approach involving using class-activation maps limits the localisation to discriminative parts of the tools, missing out significant parts of the instruments like the shafts (Fig. 2.5). Furthermore, research works on weakly-supervised learning have mostly focused on frame-wise binary tool presence annotations, which still require a certain annotation effort. This has led to overlooking the opportunity given by even cheaper sources information, like tool usage provided by robotic systems. This information describes which tools are attached to the system, without providing guarantees on their visibility in the field-of-view.

2.3.3 Weak Prior Knowledge based Solutions

Before the advent of DL, and learning-based methods in general, solving the image segmentation problem required the formalization of general knowledge about the instruments, derived from considerations of aspects like color distribution and tool positioning in the image space. We define such prior knowledge as *weak*, as opposed to the *strong* prior knowledge provided by accurate 3D models of the instruments, discussed in the next Section.
Early work by [WAH97] performs color-based segmentation using external colored markers attached to the laparoscopic instruments. Pixels belonging to the markers are selected according to a thresholding operation performed on the HSV (hue-saturation-value) color space. Spatial filtering, implemented as a convolution operation with a kernel of uniform values, is then applied to the masks to reduce the effect of scattered noise (Fig. 2.6).

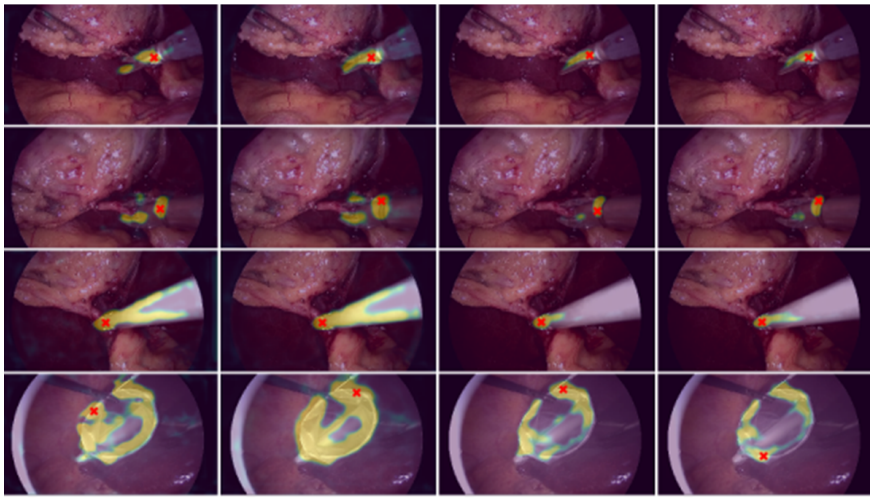

Figure 2.5: Endoscopic images overlaid with corresponding localisation maps and predicted tool centers for different weakly-supervised architectures. Note how the localisation is mostly confined to the tip of the tools. Courtesy of [VMMP18].

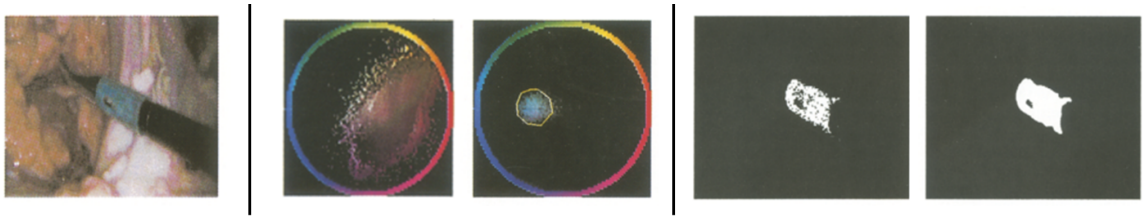

Figure 2.6: From left to right: endoscopic frame showing the colored marker applied on the laparoscopic instrument; HSV color space histogram of endoscopic frame and colored marker; segmentation masks before and after post-processing spatial filtering. Courtesy of [WAH97].

[DGDM05] improve the reliability of the HSV color-space segmentation, by combining it with an adaptive region growing algorithm with automatic seed detection. Such improvement allows to get rid of the external markers, although validation was carried out on a small-sized dataset, insufficient to assess the algorithm's robustness. During the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2006, separate works introduced the idea of exploiting constraints on shape and insertion-point of laparoscopic instruments to improve segmentation results [VLC06, DNdM06]. The first, in particular, proposes a multi-step approach for instrument segmentation, first computing potential edge points through Sobel filtering, then using them to locate instruments' symmetry axis and edges on the image space.
A seminal work by [PVH09] marked the break-through of learning-based approaches for instrument segmentation. As part of this work, aiming at instrument 3D pose estimation, and therefore detailed in Section 2.3.4, the problem of instrument segmentation was formalized as pixel-wise classification. It was solved by manually selecting a set of image features, and training a Gaussian Mixture Model to predict whether the pixel belonged or not to an instrument. Selected features included RGB and HSV color of the pixel, average intensity within a small window, and five Laplacian of Gaussian filters of

different bandwidths to add texture information. As the generalization ability of such a solution was poor, the complete approach required to have at least one manually-annotated frame for each testing sequence. Similarly, [ACO+15], [SBF14] use Random Forest algorithms to perform part-segmentation of the instruments, again as part of 3D pose estimation pipelines. As a way to boost segmentation results, without requiring manual initialization at test time, [BBO+15] propose an approach combining pixel-wise ML classification, performed by a boosted decision tree algorithm, with a tool-specific template matching, to enforce global shape consistency (Fig. 2.7).

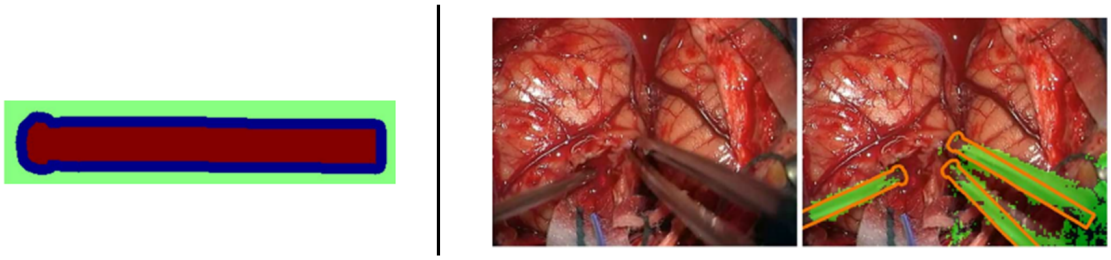

Figure 2.7: Left: Fixed shape template illustration for a suction tube used by [BBO+15]. Right: example of endoscopic frame and segmentation results of boosted decision tree algorithm (green overlay) and template matching (orange overlay). Courtesy of [BBO+15].

Following the advent of DL, manual annotations have replaced the use of prior knowledge. Manual annotations implicitly provide the same information, without requiring the effort to explicitly formalize it. However, as the interest for annotation-free approaches is rising, prior knowledge information has been recently repurposed to supervise the training of DL models. [LWJ+20a] integrate prior knowledge about color distribution and position of the instruments in the image space in a DL architecture. The approach generates segmentation pseudo-labels using such handcrafted cues, and then refines segmentation results exploiting feature correlation between adjacent video frames.

2.3.4 Strong Prior Knowledge based Solutions

3D models of surgical instruments, in combination with kinematic models and joint values, offer a *strong* and versatile source of prior knowledge to solve the surgical instrument segmentation problem. Approaches integrating parametrized 3D models of the tools commonly adopt the general framework shown in Figure 1.12. Such a framework is extremely versatile, allowing to tackle both the tasks of 3D pose estimation and image segmentation.
A seminal work by [PVH09] formalizes 3D shape estimation of robotic instruments as an iterative optimization problem aimed at estimating the pose parameters determining instruments' shape. At each step, the objective of the optimization is to align the estimated projection of surgical tools, given by the currently estimated tool configuration, with a tool part semantic segmentation mask obtained from the endoscopic frame. The segmentation mask is extracted from frames using a Gaussian Mixture Model algorithm, trained on the first few frames of each video sequence. Then, given the instrument 3D model, the calibrated camera model, and the previously estimated kinematic configuration (at the first frame the 3D model is manually aligned by the user), an optimization

algorithm is run to refine the tool configuration until the two silhouettes match, according to an overlap metric. The optimization is run using a gradient-free algorithm as the overlap metric used does not allow gradient propagation.
This approach has inspired several subsequent works. [AOH+18] in particular modify the optimization algorithm, speeding up the algorithm and improving performance (Fig. 2.8).

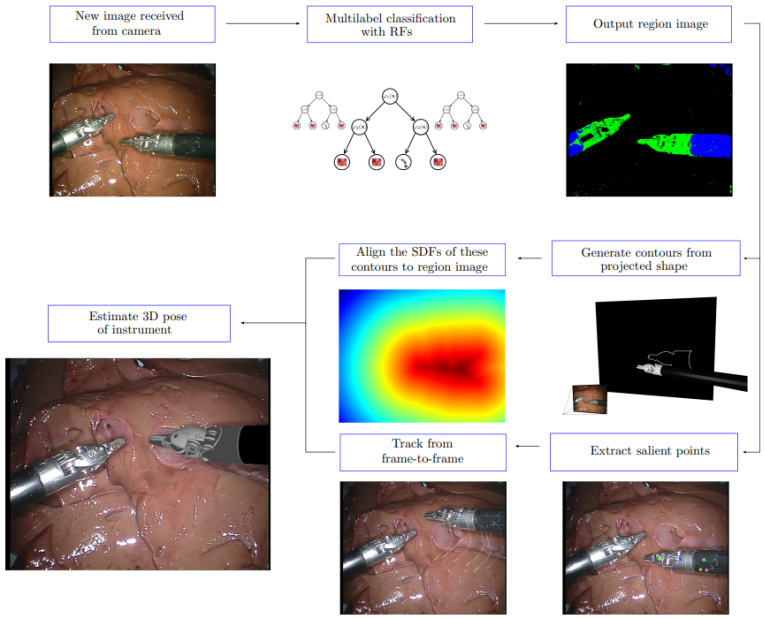

Figure 2.8: Overview of the 3D pose estimation approach by [AOH+18].

More recently, [DZK+22] have formalized CaRTS, a framework describing the causal relationship between observed kinematics and corresponding endoscopic image, refining the first based on the information provided by the latter. Interesting parallels can be drawn between [PVH09], previously discussed, and CaRTS. Both approaches use a 3D model of the instruments. Both the approaches use an image-based objective, powered by manual annotations: the first uses a contour matching loss, where the reference silhouette is obtained from a Gaussian Mixture Matrix algorithm manually initialized at the first iteration; CaRTS uses a feature matching loss, where feature representativeness is ensured by extracting them using a U-Net segmentation architecture, pre-trained for binary instrument segmentation. Finally, both initialize the optimization from a meaningful guess: the first [PVH09] initializes pose parameters manually at the first frame, then uses the previously estimated values, implicitly relying on temporal consistency; CaRTS uses observed kinematics and refine it adding an error term as optimization variable. CaRTS potential mostly relies on the scalability of the causal model proposed. In a follow-up publication, [DWLU23] CaRTS framework was extended, decoupling time-variant and time-invariant factors determining instrument configuration. This allows to separately model, and potentially optimize, kinematic values, camera-robot transformation and camera parameters. In addition, the proposed enhanced causal model also conceptually incorporates the interactions between instruments and environment, although no implementation is currently available.
Parallel works have exploited this framework to directly solve the segmentation task, without aiming at estimating the instruments' 3D pose. [dCRPR19] combine recorded

kinematic joint values and 3D kinematic models of flexible surgical instruments to generate pseudo ground truth segmentation masks. As the kinematics is often inaccurate, the obtained masks are not used to directly supervise segmentation. Instead, they are used to initialize a GrabCut segmentation algorithm. The generated pseudo-masks can be subsequently used for DL model training (Fig. 2.9). Similarly, [PSN20] generate pseudo-masks from recorded kinematics. To cope with their potential inaccuracy, they incorporate them in a cycle-GAN architecture, not explicitly requiring a direct matching between frames and pseudo-segmentation masks. [CES20] propose the combined use of recorded kinematics and green-screen recordings, in order to cheaply obtain ground truth segmentation masks for ex-vivo acquisition. In their work, kinematic data describing instrument movement are recorded during an ex-vivo experiment. A second repetition is performed reloading the recorded kinematic: this time a plain green background is used, allowing to obtain the ground truth using background subtraction techniques.

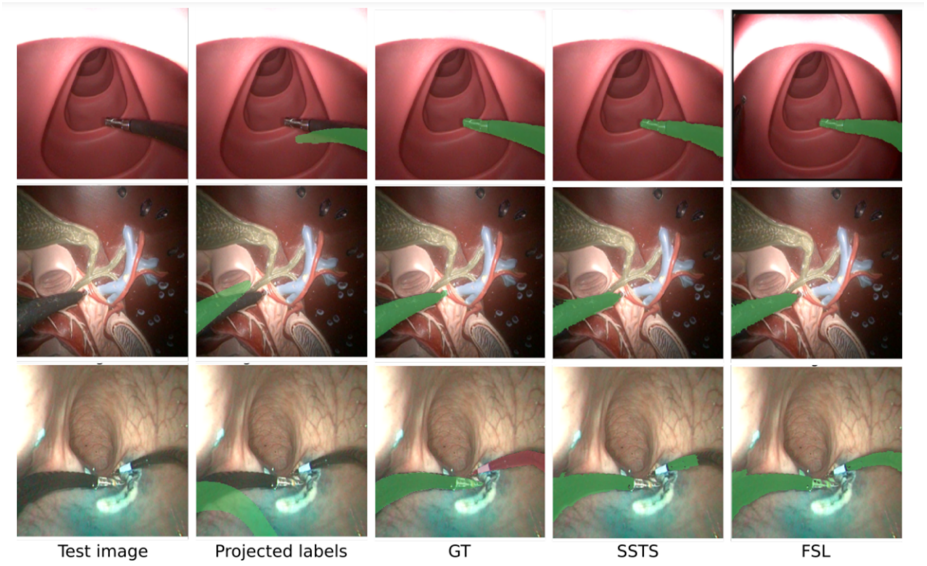

Figure 2.9: SSTS [dCRPR19] results on different datasets synthetic and in-vivo datasets. From left to right: original frame; kinematic projection (inaccurate); ground truth segmentation mask (GT); SSTS prediction; fully-supervised model prediction (FSL). Courtesy of [dCRPR19].

2.4 Thesis Positioning

Automatic localisation and identification of surgical instruments is a fundamental enabling technology for a wide range of surgical data science applications. In this Chapter we reviewed related literature on surgical instrument segmentation and 3D pose estimation, emphasizing how various sources of supervision have been explored to solve these tasks.

As discussed in Section 2.1, a large body of work has been proposed to tackle the problem of surgical instrument segmentation using a fully-supervised training formulation, requiring pixel-wise manual annotations. While effectively providing state-of-the-art

Method	Task	DL	Source of Compl./Prior Knowledge					Manual
			Simul.	Compl.		Weak PK	Strong PK	Segm.
				K	P			
[SSMZ20]	BS	✓	✓					
[SMZ21]	BS	✓	✓					
[CS21]	BS	✓	✓					✓
[GPHFD+21]	BS	✓	✓					
[MMC+21]	BS	✓	✓					✓
[VMMP18]	BB	✓			✓			
[NMMP19]	BB	✓			✓			
[XLL+22]	BB	✓			✓			
[WAH97]	BS					✓		
[DGDM05]	BS					✓		
[VLC06]	BS					✓		
[LWJ+20a]	BS	✓				✓		
[PVH09]	3D						✓	✓
[AOH+18]	3D						✓	✓
[DZK+22]	3D	✓		✓			✓	✓
[dCRPR19]	BS	✓		✓		✓	✓	
[PSN20]	BS	✓		✓			✓	
[CES20]	BS	✓		✓			✓	
FUN-SIS	BS	✓				✓		
PAF-IS	IS	✓			✓	✓		
KI-BOT	3D	✓		✓			✓	

Table 2.1: Methods presented in Section 2.3, and thesis contributions (FUN-SIS, PAF-IS, KI-BOT, last three rows), exploiting prior and complementary information to solve various instrument localisation tasks. For each solution we highlight, from left to right: the task, among binary segmentation (BS), instance segmentation (IS), bounding-box detection (BB) and 3D pose estimation (3D); if Deep Learning based (DL) or not; if based on simulation/semi-synthetic data (Simul.), complementary information (Compl.), in the form of kinematics (K) or binary tool presence information (P), weak or strong prior knowledge (PK); if requiring manual annotations, in the form of segmentation masks (Segm.), at some step of the method.

accuracy on benchmark datasets, such attempts totally rely on the availability of manual annotations. As a result, their performance drastically drops when applied to different datasets, if no labels are available for retraining [KJD+21]. As discussed in Section 1.4.2, acquiring additional manual annotations can be costly and time-consuming, and can limit effective state-of-the-art advancement.

The work presented in this thesis aims at showing possibilities to unleash the potential of unlabelled data, in combination with prior and complementary knowledge about surgical instruments. Some directions have already been explored in literature, as discussed in Section 2.3. Several of these attempts, presented in Section 2.3.1, fall under the definition of *data-synthesis* approaches, trying to generate annotated datasets either from domain translation of simulation images [SMZ21], or from purposely collected data, like green-screen recordings [GPHFD+21]. While promising, the performance and applicability of these methods is still limited by factors such as the quality of translation

and the need for ad-hoc setups to collect synthetic data.

Beyond data-synthesis, weak prior knowledge has been largely used in early works for tool segmentation [WAH97, DGDM05, VLC06], but commonly overlooked after the deep learning breakthrough in the field. Only recently, [LWJ+20a] tried to formalize general assumptions on instrument color distribution and positioning to train a deep learning model for the task of binary segmentation. However, the chosen sources of prior knowledge still require domain-specific tuning, leading to sub-optimal results compared to fully-supervised solutions. The work presented in Chapter 3 (FUN-SIS) explores the injection of a more general and robust prior knowledge of instruments' shape and motions. This knowledge is easier to formalize and is applicable to different surgical domains with no domain-specific tuning.

Complementary information about surgical tools, like binary tool presence, has been extensively explored for tool localisation via bounding-box detection [VMMP18, NMMP19]. However, the standard use of class-activation maps, limits the localisation to discriminative parts of the tools, missing out significant parts of the instruments like the shafts. Therefore a clear way to apply such knowledge for the segmentation tasks is still missing in literature. The work presented in Chapter 4 (PAF-IS) tackles the problem of instance segmentation, relying on binary tool presence information.

When available, parametrized 3D models can provide a strong prior knowledge about tools, useful to solve both the tool segmentation and 3D pose estimation tasks. In addition, such models allow to integrate recorded robot kinematics into the problem, a rich source of complementary knowledge. However, such information is often noisy, making it necessary to integrate manual annotations to solve the tasks [DZK+22]. In Chapter 5 we propose a framework (KI-BOT) allowing to perform vision-based 3D pose estimation, purely relying on the recorded kinematic data for training. Differently from existing solutions [DZK+22, AOH+18], this framework is used to perform offline training of a deep learning model which directly regresses kinematic joint values from endoscopic images. This significantly speeds up inference, as our solution does not require optimization at inference time, and removes the need to access robot kinematics inside the operating room.

Table 2.1 summarizes all the approaches using prior and complementary knowledge presented in Section 2.3, and our proposed contributions. Overall, such contributions aim at advancing the state-of-the-art along two separate directions:

- by showing that prior and complementary knowledge about surgical instruments can completely replace manual annotations to train deep learning models for the tasks of binary instrument segmentation and 3D pose estimation;
- by showing ways to apply such knowledge to the task of instance segmentation, for which alternatives to full-supervision are still missing in literature.

CHAPTER 3

FUN-SIS: a Fully Unsupervised Approach for Surgical Instrument Segmentation:

Contents

3.1 Introduction 60
3.1.1 Objective & Contributions 60
3.1.2 Learning from Motion and Noisy Labels 61
3.2 Methodology 63
3.2.1 Step 1: Unsupervised Motion Segmentation 64
3.2.2 Step 2: The Proxy Segmentation Network 66
3.2.3 Step 3: Refining Noisy Labels 67
3.2.4 Training Strategy 69
3.3 Experimental Set-up 70
3.3.1 Datasets 70
3.3.2 Artificially Corrupted Datasets 72
3.3.3 Design Choices & Training Details 73
3.4 Experiments and Results Analysis 73
3.4.1 Optical-Flow Segmentation 74
3.4.2 Single-frame Binary Tool segmentation 76
3.5 Ablation Studies and Additional Experiments 81
3.5.1 Optical-Flow Augmentation and Noise Vector Size 81
3.5.2 Proxy Network Architecture 82
3.5.3 Loss Function Coefficients (α_P, α_S) 82
3.5.4 Local IoU Parameters' Impact 83
3.5.5 Shape-Priors Quality & Quantity 85
3.5.6 Noise Properties (Unpredictability & Polarization) 86

3.5.7 Per-class IoU evaluation 89
3.5.8 Random Unlabelled Data 90
3.5.9 FUN-SIS Applicability on another Domain: Cholec80 90
3.6 Discussion and Future Work 92
3.7 Conclusion 94

3.1 Introduction

This Chapter explores the use of weak prior knowledge on instrument motion and shape to tackle the binary tool segmentation problem. The value of this Chapter for this thesis is twofold: first of all, it shows the feasibility and the effectiveness of training a deep learning model on unlabelled data, relying only on weak prior knowledge information; secondly, it enables the development of our second contribution, which relies on binary segmentation information to solve the semantically richer task of instance segmentation, with minimal additional information (Chapter 4).

3.1.1 Objective & Contributions

In this Chapter, we present FUN-SIS, a Fully-UNsupervised approach for binary Surgical Instrument Segmentation (Figure 3.1). The proposed solution allows to effectively train a binary surgical tool segmentation model on completely unlabelled endoscopic videos, solely relying on implicit motion information, in the form of optical flow, and a limited set of instrument *shape-priors*. We define *shape-priors* as binary segmentation masks of surgical tools, unpaired with the video frames and not necessarily coming from the same dataset or even surgical domain. *Shape-priors* can be obtained in convenient and various ways, such as projecting 3D virtual/CAD model of surgical instruments on the image-space, automatically segmenting green-screen recordings, or using existing annotations from existing datasets. The method is designed to extract very general knowledge from a minimal amount of such *shape-priors*, and therefore it does not require exact templates for the tools present in the unlabelled video data.
Overall, the method trains a segmentation model on pseudo-label masks generated from optical flow images; such a supervision signal, often noisy, is refined by exploiting its peculiar noise properties, stabilizing the training of the segmentation model and boosting its performance.
In order to achieve this, we make the following contributions:

- we propose a new *generative-adversarial* approach for surgical tool segmentation of optical flow images, based on simultaneous generation and segmentation of optical flow images from the *shape-priors*. Compared to common video object segmentation approaches, we relax the commonly adopted hypothesis of uncorrelated background-foreground motion, generally not verified in the surgical domain, letting the *generative-adversarial* training process adapt to the domain characteristics. This leads to state-of-the-art results both on surgical and general Video Object Segmentation datasets;

- we extensively investigate the noise properties of the segmentation masks generated using the proposed optical flow segmentation approach (*pseudo-labels*), and their impact on neural network training. We identify and thoroughly analyze two notable properties, namely *unpredictability* and *polarization*, and show that they can be exploited to largely improve segmentation results;

- we propose a novel *learning-from-noisy-labels* strategy, based on an extended *teacher-student* approach, allowing to train a *student* model only on *probably* well-labelled regions of the noisy pseudo-labels. Differently from existing approaches, usually requiring a *teacher* model trained on clean labels, we carry out an efficient region selection in a fully-unsupervised way, exploiting the aforementioned noise properties. The proposed approach leads to high-quality binary segmentation results on several surgical datasets, including the popular EndoVis 2017 Instrument Segmentation dataset, while being trained on completely unlabelled videos.

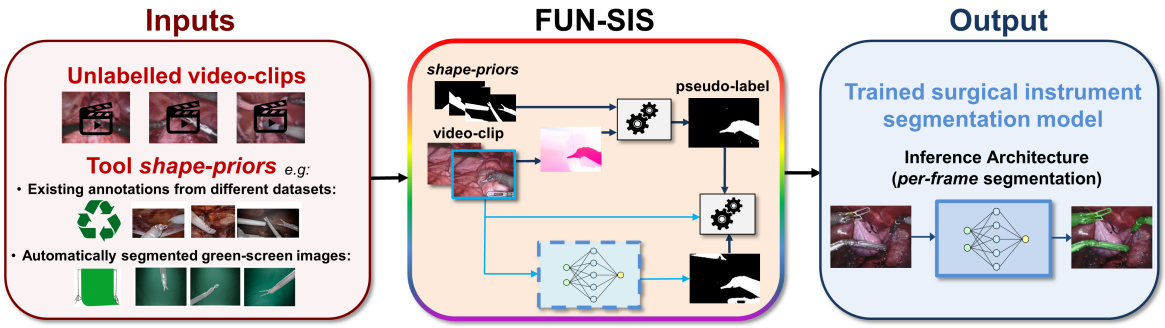

Figure 3.1: Chapter contribution from the input-output point-of-view. The proposed FUN-SIS approach allows to train a model for surgical tool segmentation requiring as inputs only unlabelled video-clips and tool shape-priors, obtainable in various convenient ways (e.g. by recycling existing annotations from other datasets). The method is based on a novel approach for unsupervised surgical tool segmentation of optical flow images, generating pseudo-label masks, and a newly designed learning-from-noisy-labels strategy, allowing to extract a clean supervision signal to train a single-frame binary segmentation model.

3.1.2 Learning from Motion and Noisy Labels

As discussed above, the proposed solution is tightly related to the topics of video object segmentation, for pseudo-label generation, and learning-from-noisy-labels, to effectively learn from such pseudo-labels. An overview of these topics is now presented.

3.1.2.1 Video Object Segmentation

Motion information is used by the human visual system for *perceptual grouping*, the process of organizing the visual information in order to efficiently perceive and interact with the world. In the general object segmentation context, as well as for surgical tool segmentation, motion can be a very discriminative cue, easy to obtain from unlabelled videos by means of readily available optical flow estimators like RAFT [TD20]. Given the relevance of motion, the computer vision community has been constantly exploring the task of Video Object Segmentation (VOS). The two standard approaches to it are semi-supervised VOS and unsupervised VOS. Semi-supervised VOS aims at tracking a target,

specified in the first frame of a video sequence in the form of a segmentation mask, across the following frames. In the surgical context, such an approach is not applicable, as the repeated changes of instruments during a procedure, and their motion *in* and *out* of the field of view, would require a continuous re-identification of the objects to be tracked. Unsupervised VOS, instead, aims at separating a salient foreground object from the background, based on motion information. It is worth noticing that, despite its name, unsupervised VOS has often been tackled in literature by means of fully-supervised training (e.g. [MAO+20]): the *unsupervised* attribute indicates, instead, that this family of methods does not need an initial mask of the object, as opposed to semi-supervised VOS. Unsupervised VOS approaches not requiring a ground truth supervision signal to train, commonly rely on the strong assumption of incoherent background motion, uncorrelated with foreground motion [WSYP17, YLSS19a, YLL+21]. This hypothesis is commonly not applicable to surgical scenes: foreground (tools) and background (anatomical structures) strongly interact with each other, resulting in correlated motion of the two and coherent motion of the anatomical structures (Figure 3.2).

In this work we propose a novel unsupervised approach for optical flow tool segmentation, not requiring ground truth annotations of the training data. In order to tackle the above-mentioned challenges, we relax the hypothesis of incoherent background motion, letting a generative-adversarial training process adapt to the domain characteristics.

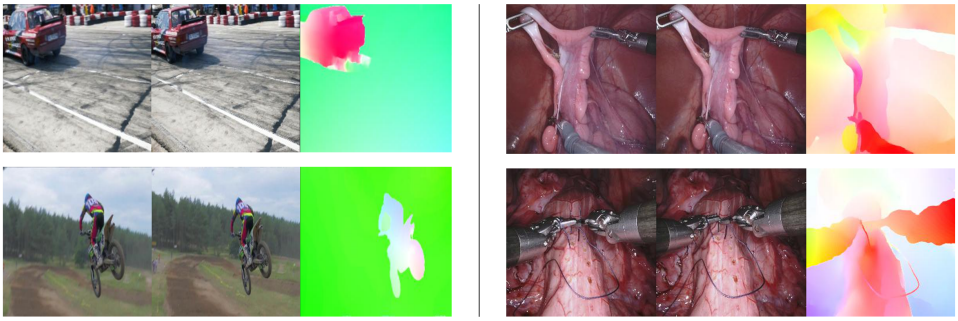

Figure 3.2: Frames and corresponding optical flow images from DAVIS dataset [PPTM+16] (left) and EndoVis 2017 dataset [ASK+19] (right). Note how, in surgical images, background motion is often coherent and correlated with foreground motion.

3.1.2.2 Learning-from-noisy-labels

Effectively learning from noisy labels is becoming an essential need of deep learning applications. In order to gather the massive amounts of annotations required to train deep learning models, researchers have recently been looking for alternatives to standard *in-house* annotation, such as crowd-sourcing [YDDM18] or automatic-labelling [GZH+16]. However, while dramatically cutting down the cost of annotations, these approaches tend to provide noisy labels. In order to tackle the *learning-from-noisy-labels* problem, several approaches have been proposed in literature, such as noise adaptation layers [CG15], robust loss designs [ZS18, WMC+19] and different strategies aimed at automatically selecting well-labelled samples [JZL+18, HYY+18]. While well theoretically motivated, the effectiveness of the above-mentioned methods has been mainly proven for the classification task in less challenging datasets compared to the surgical ones, such as artificially

modified versions of benchmark datasets like CIFAR [LeC98], and, less frequently, in real-world datasets with modest amount of noise like WebVision [LWL+17] and Clothing 1M [XXY+15].
Segmentation differs from standard classification because individual semantic labels are not independent, as they come grouped in images. This creates the need to rethink standard methods such as *sample-selection*, since discarding full samples may represent a waste of useful information, if the noise is localised in certain image regions only. Confidence map estimators proposed to tackle this challenge, commonly rely on small sets of annotated data [YWL+19, NGWS18].
In this work, we tackle the problem of learning binary surgical tool segmentation from noisy pseudo-labels obtained from unsupervised segmentation of optical flow images. Differently from the above-mentioned works, our method does not require any set of clean labels in order to perform well-labelled pixel selection from the noisy pseudo-label masks. Instead, it leverages their peculiar properties, and the favorable behaviour of neural networks when dealing with such type of noise.

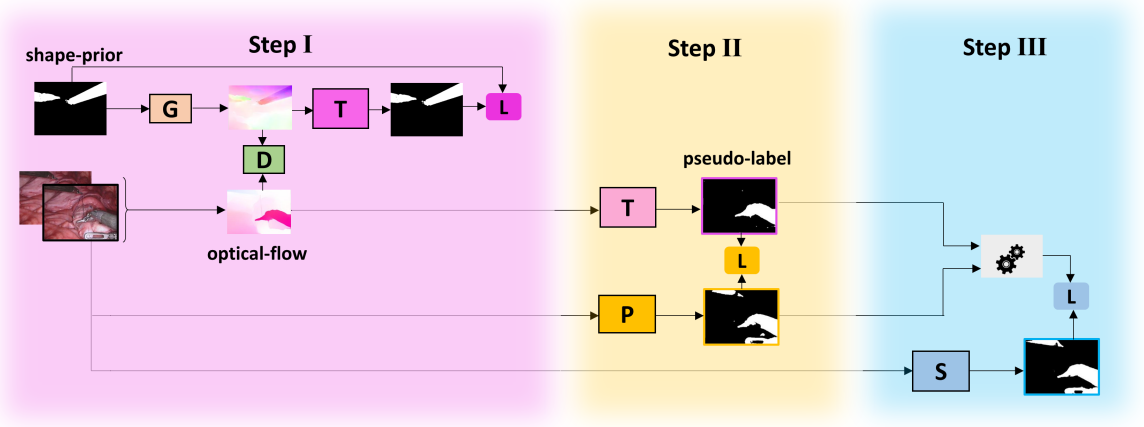

Figure 3.3: General overview of proposed FUN-SIS training architecture. Step I: training of the optical flow segmentation network (Teacher, T), as part of a generative-adversarial architecture mapping shape-priors into synthetic optical flow images and vice-versa. Step II: Proxy segmentation model training, directly supervised by the pseudo-labels obtained from optical flow segmentation by the Teacher model. Step III: Student segmentation model training on the refined supervision signal obtained by the combination of Proxy predictions and pseudo-labels. Main training losses (L) are also shown.

3.2 Methodology

The FUN-SIS approach (Figure 3.3) is a 3-step method that carries out unsupervised surgical tool segmentation of optical flow images (step I) and subsequently trains a single-frame binary segmentation model on the noisy pseudo-labels generated at step I using a *learning-from-noisy-labels* strategy to refine the supervision signal (steps II and III). The 3 steps are introduced below and detailed in the next sections:

I) generative-adversarial training of the optical flow tool segmentation model (called *Teacher*), carried out by simultaneously learning to generate and segment synthetic optical flow images from tool *shape-priors* (Section 3.2.1, Figure 3.3-I);

II) training of a model (called *Proxy*) for tool segmentation of individual frames, using, as direct supervision, the noisy pseudo-labels generated by the *Teacher* model via optical flow segmentation; the effectiveness of this step is guaranteed by a property of the noise affecting the pseudo-labels, called *unpredictability* (Section 3.2.2, Figure 3.3-II);

III) training of a model (called *Student*) for tool segmentation of individual frames, using, as supervision, only *probably* well-labelled regions of the pseudo-labels, selected according to the local agreement between the *Teacher* and *Proxy* models; the effectiveness of this step is guaranteed by another property of the noise affecting the pseudo-labels, called *polarization* (Section 3.2.3, Figure 3.3-III).

3.2.1 Step 1: Unsupervised Motion Segmentation

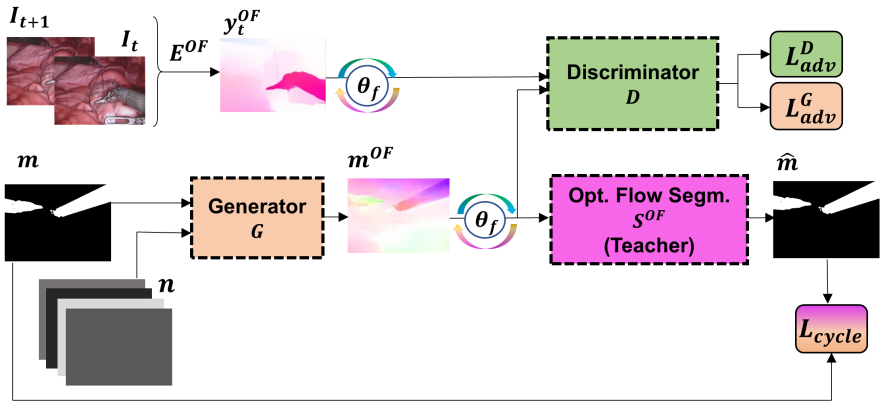

Figure 3.4: Overview Step I of FUN-SIS: generative-adversarial training of optical flow segmentation model S^{OF} (Teacher), generator (G) and discriminator (D); generated (m^{OF}) and real ($E^{OF}(I_t, I_{t+1})$) optical flow images undergo augmentation via random rotation θ_f. A noise vector $\mathbf{n}$ *is concatenated to the shape-prior m to allow one-to-many mapping. Loss boxes (L) are color coded to show which models are responsible for their minimization during training.*

The proposed approach for unsupervised optical flow-segmentation is based on a generative-adversarial approach, constrained by a cycle-consistency loss. This approach allows to learn the mapping between the domain of optical flow images and the domain of *shape-priors*, consisting of realistic binary segmentation masks of the target object (in this case surgical tools), without requiring pairwise matching between the two domains. The method is inspired by the classic cycle-GAN architecture [ZPIE17], a popular generative architecture for image-to-image translation from unpaired domains. However, it is known that mapping between a domain of minimal complexity, as the binary *shape-priors*, lacking strong discriminative features, and a more complex one, such as the optical flow, is an ill-posed problem, suffering from issues such as information-hiding (*'steganography'* [CZS17]) and overpowering discriminator, possibly hindering the whole training process.
In order to deal with this *complexity-imbalance*, we propose the following modifications to the standard cycle-GAN:

- we use a single cycle-consistency loss (only for *shape-priors* domain), in order to avoid reconstructing a high-complexity domain sample from a *synthetic* low-complexity domain sample, preventing *'steganography'*;

- we concatenate the *shape-priors* domain samples with a random noise vector before feeding them to the generator. This allows the generator to produce different *synthetic* optical flow images from the same *shape-priors* mask, disentangling the tool silhouette from its motion;

- we make intensive use of on-the-fly image augmentation.

The architecture for the proposed optical flow segmenter is displayed in Figure 3.3-I, and discussed below.
Let us consider two consecutive frames belonging to a video, I_t, I_{t+1} (original frames augmented by an augmentation protocol $AugmData$, consisting of random cropping and flipping), an optical flow estimator $E^{OF} : \{I_t, I_{t+1}\} \rightarrow y_t^{OF}$, where y_t^{OF} is the optical flow image in the form of $[u, v]$ pixel displacement, an optical flow generator model G, an optical flow segmentation model S^{OF} (also referred to as *Teacher* model, due to its role in steps II and III), a *shape-priors* binary mask m and a discriminator model D. The generator G takes as input the *shape-priors* mask m, augmented on-the-fly by an augmentation protocol $AugmMask$, consisting of random cropping and flipping, and concatenated with a noise vector $\mathbf{n}$, sampled from a normal distribution of mean μ and standard-deviation σ, and resized to the input mask resolution, and outputs a synthetic optical flow image m^{OF}, also in the form of $[u, v]$ pixel displacement. Both the real and synthetic optical flow images, y_t^{OF} and m^{OF}, undergo on-the-fly augmentation, based on augmentation protocol $AugmFlow$, and following normalization operations:

- ***AugmFlow***: the optical flow is multiplied by a random rotation matrix in the form:

$$R = \begin{bmatrix} \cos\theta_{flow} & -\sin\theta_{flow} \\ \sin\theta_{flow} & \cos\theta_{flow} \end{bmatrix}, \tag{3.1}$$

  where θ_{flow} is randomly picked from a uniform distribution. This operation, performed on-the-fly, increases the variability of the optical flow, and releases the generator from the burden to generate every possible flow direction;

- **normalization:** each optical flow image is normalized by dividing it by the maximum pixel displacement $\sqrt{u^2 + v^2}$ in it. This operation keeps the generated optical flow image in a controlled range (where maximum displacement has norm equal to 1).

The synthetic optical flow image m^{OF} is then fed to the optical flow segmentation model S^{OF}, which outputs the *cycled shape-priors* mask $\hat{m}$. The real and synthetic optical flows y_t^{OF} and m^{OF} (both augmented and normalized) are fed to the discriminator D, which is trained to distinguish among the two. Cycle-consistency is ensured by requiring the cycled-mask $\hat{m}$ to match the input mask m by means of a standard cross-entropy loss:

$$L_{cycle} = -m\log(\hat{m}) - (1-m)\log(1-\hat{m}). \tag{3.2}$$

Discriminator's outputs are used to enforce realistic appearance of m^{OF} by training the discriminator D and the optical flow generator G in an adversarial way. Specifically, the adversarial loss functions are defined as:

$$L_{adv}^{G} = -\log(D(m^{OF})), \tag{3.3}$$

$$L_{adv}^{D} = -\log(1 - D(m^{OF})) - \log(D(y_t^{OF})). \tag{3.4}$$

The full architecture is trained end-to-end. The discriminator D is trained to minimize L_{adv}^{D}, the optical flow segmenter S^{OF} is trained to minimize L_{cycle}, the optical flow generator G is trained to minimize the sum of L_{adv}^{G} and L_{cycle} :

$$L^{G} = L_{adv}^{G} + L_{cycle}. \tag{3.5}$$

3.2.2 Step 2: The Proxy Segmentation Network

The optical flow segmentation by S^{OF} (*Teacher* model) is used to generate pseudo-labels for the unlabelled frames: each frame I_t is paired with the *Teacher*-generated pseudo-label mask $y_t^T = S^{OF}(y_t^{OF})$, which is used as direct supervision to train a neural network (*Proxy* model) to perform tool segmentation of individual frames (Figure 3.3-II).

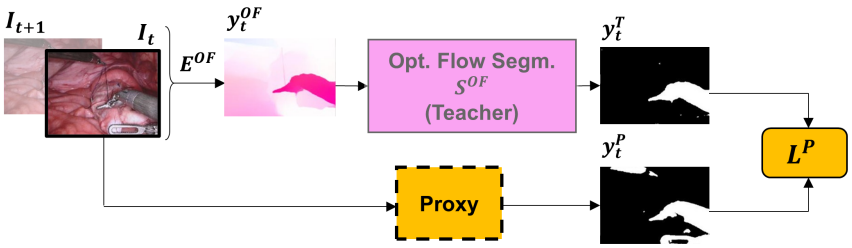

Figure 3.5: Overview of Step II of FUN-SIS: Proxy segmentation model training, directly supervised by the pseudo-labels y_t^T, obtained from optical flow segmentation by the Teacher model. Loss boxes (L) are color-coded to show which models are responsible for their minimization during training.

The proposed approach to leverage the noisy pseudo-labels relies on findings from [AJB+17], which show that, while neural networks are in principle capable of memorizing noisy samples, they tend to first take advantage of shared patterns across training examples, given their finite capacity. In a parallel study, [RVBS17] empirically confirmed, in the classification task, that neural networks can generalize well even when trained on massively noisy data, rather than just memorizing noise, assuming that the label noise is not conditioned by the corresponding input image itself. We define this condition as the ***unpredictability*** property.
The noise affecting the pseudo-labels y_t^T can be divided into two additive processes: the optical flow estimation noise and the optical flow segmentation noise. In both cases, the property of *unpredictability* of noise affecting the pseudo-label y_t^T, from the single frame I_t, holds:

- the possible absence of tool motion or presence of background coherent motion in the optical flow image $y_t^{OF} = E^{OF}(I_t, I_{t+1})$, potential sources of y_t^T noise, cannot be

predicted from the individual frame I_t only, but requires an additional frame (I_{t+1}) to be predicted;

- the optical flow segmentation used to generate the pseudo-labels ($y_t^T = S^{OF}(y_t^{OF})$), a second possible source of noise due to the inevitable sub-optimality of S^{OF} model, does not involve the use of the frame I_t, contrarily to standard VOS approaches, where both frame and optical flow are used to make a prediction (e.g. [YLSS19a]).

Given the *unpredictability* property, we can train a neural network (*Proxy* model) to perform single-frame tool segmentation, using the noisy pseudo-labels y_t^T directly as supervision signal. The *Proxy* network takes as input the frame I_t and outputs the segmentation mask y_t^P. The network is trained to minimize the loss L^P, which is the sum of the binary cross-entropy loss L_{CE}^P and the log Intersection-over-Union loss L_{IoU}^P, weighted by a factor α_P:

$$L_{CE}^P = -y_t^T \log(y_t^P) - (1 - y_t^T) \log(1 - y_t^P), \tag{3.6}$$

$$L_{IoU}^P = -\log \frac{\sum(y_t^P y_t^T)}{\sum(y_t^P + y_t^T - y_t^P y_t^T)}, \tag{3.7}$$

$$L^P = \alpha_P L_{IoU}^P + (1 - \alpha_P) L_{CE}^P. \tag{3.8}$$

During training, the *Proxy* network, unable to learn the noisy pattern from the pseudo-labels, tries to fit them with the *easiest* compatible pattern, experimentally shown to be the separation of tools from anatomy. In order to encourage this effect, we suggest the advantage of using a relatively small-capacity network compared to deeper ones. In fact, in principle, a neural network of infinite capacity would be able to memorize each training sample as a look-up table. The use of a small-capacity network forces the model to find a common pattern to fit the data, reducing the chances memorize the noisy labels. We experimentally investigate this aspect in our ablation studies, reported in Section 3.5.2. However, as the training progresses and the pattern is learnt, the loss does not get further minimized, and gradient descent updates remain high. This prevents convergence to an optimal solution, which mainly affects *Proxy* segmentation accuracy on hard-to-classify pixels, such as the boundary ones. This shortcoming is addressed and mitigated at step III below.

3.2.3 Step 3: Refining Noisy Labels

Together with the *unpredictability* property, a second peculiar property of the noise affecting the pseudo-labels y_t^T derives from the fact that individual tools, moving coherently, tend to have a uniform appearance in the optical flow image; this implies that, under ideal conditions (optimal optical flow estimator E^{OF}, optimal optical flow tool segmenter S^{OF}), each individual tool will be either perfectly segmented (if moving) or completely mislabelled (if not moving). We define the resulting noise feature as ***polarization*** property, as a tool can ideally only be perfectly segmented or completely mislabelled by optical flow segmentation. In the real case, this property still holds, although occlusions and sub-optimal optical flow estimation/segmentation tend to

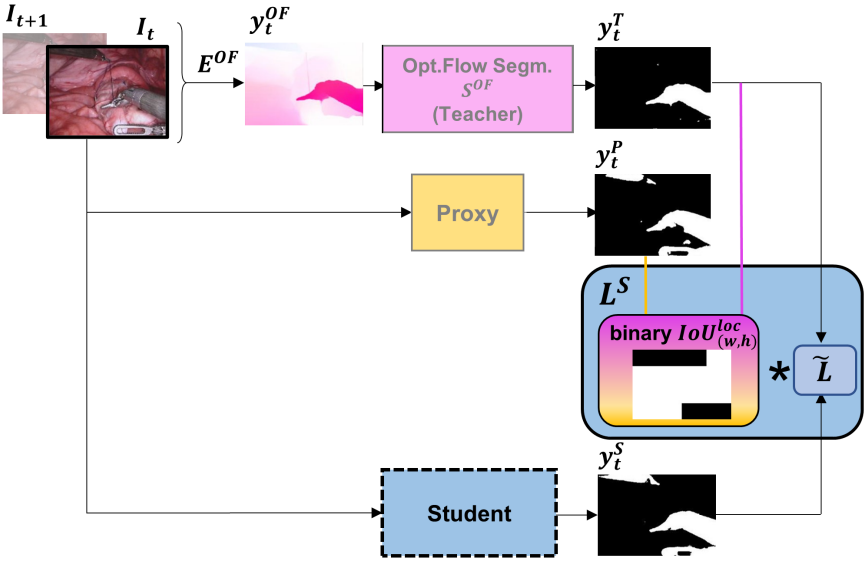

Figure 3.6: Overview of Step III of FUN-SIS: Student segmentation model training, leveraging local Intersection-over-Union ($IoU^{loc}_{(w,h)}$) between Teacher and Proxy predictions to select well-labelled regions of y_t^T. $\tilde{L}$ is a pixel-wise loss (e.g. cross-entropy), masked by the pixel-wise multiplication ($*$) with the binarized local IoU. Loss boxes (L) are color-coded to show which models are responsible for their minimization during training.

inevitably reduce the intensity of the *polarization* (i.e. there will possibly be partially segmented tools). As a practical corollary, the *polarization* property suggests that inside a pseudo-label y_t^T, there will be either almost-perfectly labelled or almost-completely wrongly-labelled regions. This *polarization* property will be thoroughly investigated in the experiments from Section 3.5.6.

In order to improve training robustness and consistency, we exploit the *polarization* property by designing an unsupervised method to select well-labelled regions of the pseudo-labels y_t^T (Figure 3.3-III). The criterion adopted for this selection is the agreement between *Proxy* network predictions y_t^P (binarized using a threshold value ϵ_P), and pseudo-labels y_t^T (binarized using a threshold value ϵ_T). The underlying idea is that the *Proxy* network learns a robust general representation (the *easiest* pattern). While its predictions can be incorrect at small-scale (e.g. on border pixels), they are overall reliable at greater scale (i.e. tools are not completely mislabelled as possibly happening in the pseudo-labels). In order to leverage this observation, we introduce a local version of the Intersection-over-Union (IoU) metric, called ***local* IoU** ($IoU^{loc}_{(w,h)}$). In order to compute $IoU^{loc}_{(w,h)}$ between two masks, a window of size $w \times h$ is slid across the masks, using a stride equal to the window size, and IoU is computed inside each time. The output is an image with the same resolution as the input masks, whose value at each pixel is the IoU computed for the region containing the pixel (Figure 3.7). Due to the way it is constructed, it holds that:

$$\frac{1}{W \cdot H} \sum IoU^{loc}_{(W,H)} = IoU, \tag{3.9}$$

$$\frac{1}{W \cdot H} \sum IoU^{loc}_{(1,1)} = PA, \tag{3.10}$$

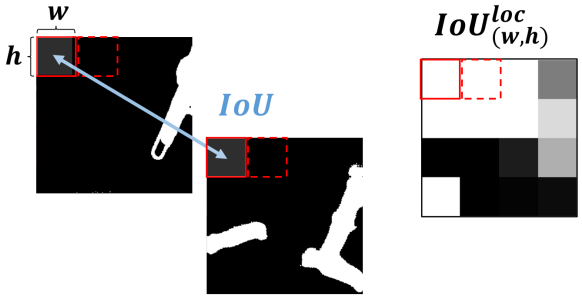

Figure 3.7: Local IoU $IoU^{loc}_{(w,h)}$ is computed by sliding a window of size $w \times h$ on the two input masks, computing standard IoU at each corresponding location. The output is a single-channel image, having the same resolution as the input masks, with each pixel's value being set to the one of the IoU computed for the region it belongs to.

where $W \times H$ is the size of the input masks, PA is the pixel accuracy metric and the summation is performed over pixels. This makes *local* IoU a metric that interpolates between standard IoU and pixel accuracy, by varying the window size parameter. *Local* IoU is computed between pseudo-label y^T_t and *Proxy* prediction y^P_t, and then binarized using a threshold parameter ϵ_{IoU}. ϵ_{IoU} represents the minimum agreement between *Proxy* and *Teacher* required for a region of y^T_t to be regarded as well-labelled. The binarized *local* IoU, $\overline{IoU}^{loc}_{(w,h)} = bin(IoU^{loc}_{(w,h)}, \epsilon_{IoU})$ is used to prevent the loss propagation through the *probably* wrongly-labelled regions of the pseudo-labels y^T_t, during the training of the *Student* network. In particular, the *Student* network takes as input the frame I_t and outputs the segmentation mask y^S_t. The network is trained to minimize the loss L^S, which is the weighted sum of binary cross-entropy loss L^S_{CE} and log Intersection-over-Union loss L^S_{IoU}, masked by multiplying each pixel-wise loss by $\overline{IoU}^{loc}_{(w,h)}$:

$$L^S_{CE} = \frac{1}{\sum \overline{IoU}^{loc}_{(w,h)}} \overline{IoU}^{loc}_{(w,h)} (-y^T_t \log(y^S_t) - (1 - y^T_t) \log(1 - y^S_t)), \tag{3.11}$$

$$L^S_{IoU} = -\frac{1}{\sum \overline{IoU}^{loc}_{(w,h)}} \log \frac{\sum (y^S_t y^T_t \overline{IoU}^{loc}_{(w,h)})}{\sum (y^S_t + y^T_t - y^S_t y^T_t) \overline{IoU}^{loc}_{(w,h)}}, \tag{3.12}$$

$$L^S = \alpha_S L^S_{IoU} + (1 - \alpha_S) L^S_{CE}. \tag{3.13}$$

Multiplying the pixel-wise segmentation losses by $\overline{IoU}^{loc}_{(w,h)}$ allows to prevent *Student* network training on both potential false-positive regions (background regions incorrectly segmented by the *Teacher* network) and false-negative regions (tools incorrectly considered as background by the *Teacher*).

3.2.4 Training Strategy

As presented in Section 3.2 and shown in Figure 3.3, the proposed approach involves a 3-step training, where the *Teacher*, *Proxy* and *Student* models are trained successively. However, relying on the hypothesis that a neural network will not be able to memorize noisy labels, discussed in Section 3.2.2, we suggest that the *Proxy* network can be trained on the pseudo-labels produced by *Teacher* network while the *Teacher* network is being

trained. This allows the training to be a more compact, 2-step process, with steps I and II carried out simultaneously. Comparison between 3-step and 2-step training is reported in Section 3.4.2.

3.3 Experimental Set-up

3.3.1 Datasets

In order to validate the proposed contributions, extensive experiments were carried out, both on surgical and general object segmentation datasets. All the data used in our experiments are now presented and categorized as *Video* and *Shape-priors*. Details about their use in the experiments are also reported.

Video data:

- **EndoVis2017** [ASK+19]: dataset from the 2017 MICCAI EndoVis Robotic Instrument Segmentation Challenge. The dataset contains 10 video clips of abdominal porcine procedures, performed using da Vinci Xi systems. Each video contains a total of 300 high-resolution frames (1280×1024), recorded at 2 Hz. In the challenge, 8x 225 frames were used for training, while the remaining 8x 75 frames and another 2x 300 frames were held out by the organizers for testing. According to the challenge rules, man-made tools not belonging to the da Vinci system (e.g. drop-in Ultra-Sound probe), labelled by the organizers as part of a class called *Other*, are to be included in the *background* class for the binary segmentation task. This introduces the need for a model to perform a semantic differentiation inside the *instrument* class (da Vinci instruments and *Other* instruments), which goes beyond the scope of motion-based segmenters. For this reason, we refer to the dataset labelled according to the challenge rules as **EndoVis2017Challenge**, and also consider a second version of it, called **EndoVis2017VOS**, where both da Vinci and other man-made tools are labelled as *instrument*. Manually annotated masks for these tools are present in the original challenge dataset as part of the set of semantic segmentation annotations, and were combined with the available binary segmentation masks. For the main experiments, we report results on both **EndoVis2017Challenge** and **EndoVis2017VOS**. We provide results on this dataset according to 2 modalities: 1) following the same evaluation protocol as [SRKI18a], by performing 4-fold cross-validation on the 8x 225 released training data (regrouped in 4 splits), and reporting the average metric on the 4 splits, for a direct and fair comparison with other state-of-the-art approaches; 2) by training on RandSurg, a dataset of unlabelled data, described below, and testing on the 8x 225 EndoVis2017VOS frames.

- **RandSurg**: this dataset consists of 4 full unlabelled laparoscopic robotic-assisted procedures downloaded from a public repository [Wor]: adhesiolysis (1036 frames), inguinal hernia repair (1075 frames), appendectomy (500 frames) and ex-vivo suturing demo (525 frames). A set of experiments was carried out by training our model on this dataset and evaluating the performance on EndoVis2017VOS; in order to simulate a realistic application of the FUN-SIS method, and show its ease-

of-use, the videos underwent minimal pre-processing (cropping, no trimming, so possibly including out-of-body scenes).

- **STRAS**: this dataset is obtained from endoscopic submucosal dissection procedures performed through the STRAS robotic system [DDZZ+13], a robotic system consisting of a robotized endoscope, having two lateral channels for flexible robotic tools. The dataset was built from a 5 day-experiment on porcine models[1] [ZNZ+17], recorded at 30 fps. Each frame was paired with another 1 second apart in the future, for optical flow computation. The whole dataset was resampled regularly, yielding a total of 5644 frames (~1100 per experiment day). For each day, 200 frames, regularly spaced, were manually annotated for evaluation (1000 annotated samples in total). The dataset contains challenging sequences, involving bleeding, smoke, strong tool-tissue interaction and image blurring. We provide results on this dataset by performing 5-fold cross-validation (each fold corresponding to an experiment day), and reporting the average metric on the 5 splits.

- **Cholec80** [TSM+16]: dataset containing 80 unlabelled videos of manual laparoscopic cholecystectomy procedures captured at 25 Hz and resampled at 1 Hz. We provide qualitative results on this dataset by using the standard split (40 videos for training, 40 videos for testing) to show cross-surgery applicability of the proposed FUN-SIS method.

- **DAVIS2016** [PPTM+16]: a popular VOS dataset, containing different moving objects (e.g. animals, people, cars). The dataset consists of 50 clips for a total of 3455 1080p frames with pixel-wise annotations. We provide results on this dataset in order to evaluate the proposed optical flow segmentation approach on non-surgical videos. To this aim, the standard training-test split was used (30 videos for training and 20 for testing), for a fair comparison with state-of-the-art VOS approaches.

Shape-priors:

- **RoboTool**: 514 manually segmented tool masks, from the RoboTool dataset, released by [GPHFD+21]. Examples of the original frames and manually segmented tools can be seen in Figure 3.8, bottom. Original masks were cropped to remove the lateral black bands, and resized to 256×256 regardless of their original aspect ratio.

- **GrScreenTool**: automatically segmented tools from recordings in front of a green-screen. A total number of 1100 masks were downloaded from the publicly released dataset by [GPHFD+21], mostly having a single tool. Random couples of masks were then selected and merged together, in order to avoid having single-tool masks. Following this strategy, a total number 2200 masks were obtained. Examples of the original green-screen images and extracted tools can be seen in Figure 3.8, top.

- **STRASMasks**: 2000 projections of approximate 3D virtual/CAD model of the two STRAS tools, used as *shape-priors* in the STRAS experiments; details regarding the projection operation can be found in [SRDM+21].

[1] The study protocol for this experiment was approved by the Institutional Ethical Committee on Animal Experimentation (ICOMETH No.38.2011.01.018). Animals were managed in accordance with French laws for animal use and care as well as with the European Community Council directive no. 2010/63/EU

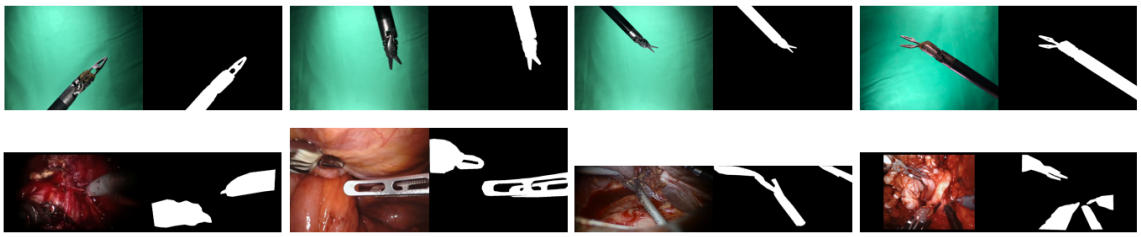

Figure 3.8: Examples of shape-priors used for the EndoVis2017 experiments, and corresponding source image. Top: tools recorded in front of the green-screen and automatically segmented [GPHFD+21], called GrScreenTool; Bottom: frames from multiple robotic-assisted laparoscopic surgeries, manually segmented as part of the RoboTool dataset [GPHFD+21]. Frames (and also masks) in this dataset come with various resolution/aspect ratios. Note how the appearance of the two domains is different: this is mainly due to the fact that GrScreenTool dataset, recorded using an external camera, show a different point of view on the instruments with respect to the standard surgical camera.

- **SegTrackV2** [LKH+13]: 976 manual annotations from the generic VOS dataset SegTrackV2. The dataset includes different segmented objects (e.g. animals, cars, people), used as *shape-priors* in the DAVIS2016 experiments.

- **FBMS59** [OMB13]: 720 manual annotations from the generic VOS dataset FBMS59. The dataset includes different segmented objects (e.g. animals, cars, people), used as *shape-priors* in the DAVIS2016 experiments.

3.3.2 Artificially Corrupted Datasets

In order to gain a full understanding of the impact of the noise properties presented in Section 3.2.2 and 3.2.3 on the proposed *learning-from-noisy-labels* approach, we also perform experiments on the EndoVis2017VOS dataset under controlled noise conditions. For these experiments we substitute, in our training pipeline, the pseudo-labels y_t^T generated by the *Teacher* network, with artificially corrupted versions of the clean labels. To this aim, we consider three types of label corruption, described below:

- *Systematic-Erosion*: each ground truth mask is eroded;

- *Erosion & Dilation*: each ground truth mask is randomly eroded or dilated;

- *Tool-Drop*: *full* tool annotations are randomly dropped (i.e. each tool is either *perfectly*-annotated or not-annotated at all).

For each noise type we apply the corresponding transformation, modulating its intensity in order to obtain 4 datasets, {D80, D60, D40, D20}, each one having a mean IoU between the corrupted labels and the ground truth of ~80%, ~60%, ~40%, ~20%, respectively (e.g. greater erosion is applied to generate D20 compared to D40, in the *Systematic-Erosion* experiment). Examples of the datasets are shown in Figure 3.9. We use this dataset as part of the ablation study detailed in Section 3.5.6.

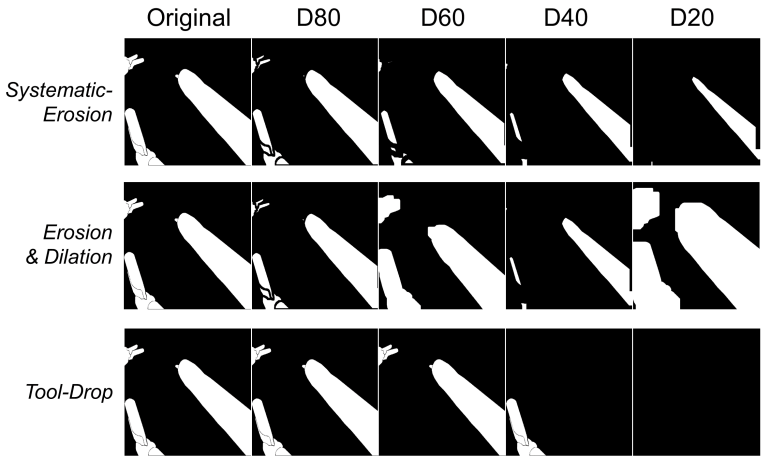

Figure 3.9: Samples from the artificially-corrupted versions of EndoVis2017 dataset. From top to bottom: Systematic Erosion, Erosion & Dilation, Tool-Drop. For each noise source, a sample from D80 (~80% mean IoU between training sample labels and original ones), D60 (~60% mean IoU), D40 (~40% mean IoU), D20 (~20% mean IoU) is shown.

3.3.3 Design Choices & Training Details

All models are implemented as neural networks. Neural network architectures and hyper-parameters were determined from preliminary experiments on external data (*phantom* dataset from [SRDM+21]), and can be found in [SRDM+23]. All the segmentation models have a U-Net-like architecture. The *Proxy* and *Student* networks have slightly different architectures, with the *Proxy* having a 11-convolutional-layer encoder (which we refer to as Unet11) and the *Student* a 16-convolutional-layer (Unet16). Optical flow estimation was carried out using RAFT [TD20], a state-of-the-art approach, trained on the publicly available non-surgical dataset FlyingThings [MIH+16]. Training and evaluation were all carried out on 256×256 resized versions of the images, regardless of their original resolution/aspect ratio, due to memory constraints. The size of the noise vector $\mathbf{n}$ was set to 32, and investigated in Section 3.5.1. Each value of $\mathbf{n}$ was drawn from a normal distribution of mean μ equal to 0 and standard-deviation σ equal to 1. The $IoU^{loc}_{(w,h)}$ window size $w \times h$ was set to 64×64 (1/4 of the image size); the threshold ϵ_{IoU} was set to 0.5. An in-depth study regarding w and ϵ_{IoU} was carried out and reported in Section 3.5.4. The loss balancing factors α_P, α_S from Equations 3.8&3.13 were set to 0.8, and investigated in Section 3.5.3. Augmentations $AugmMask$ and $AugmData$ were implemented by applying random left-right, up-down flipping and random cropping, with minimal cropped region size equal to 224×224, then bilinearly resampled to 256×256. The angle θ_{flow} for the flow rotation in $AugmFlow$ was randomly picked in the range $[-\pi, \pi]$. All augmentations were applied on-the-fly. Training was carried out using a single NVIDIA Tesla V100 GPU (32 GB).

3.4 Experiments and Results Analysis

In this section we present experimental results and comparisons with state-of-the-art methods. First, we analyze the effectiveness of the proposed optical flow segmentation approach, both on surgical and general object-segmentation datasets. We then analyze

	Annot. [%]	EndoVis2017	DAVIS2016
$\text{Baseline}_{\text{FS}}$	100	60.47±27.03	73.58±22.31
CIS [YLSS19a]	0*	24.15±21.63	60.89 (71.5)
$\text{Teacher}_{\text{RoboTool}}$(ours)	0	40.08±26.70	/
$\text{Teacher}_{\text{GrScreenTool}}$(ours)	0	**40.47±25.62**	/
$\text{Teacher}_{\text{FBMS}}$(ours)	0	/	62.72±19.23
$\text{Teacher}_{\text{SegTrackV2}}$(ours)	0	/	**63.40±19.35**

Table 3.1: Optical flow segmentation. Comparison of the proposed method (Teacher), using different shape-priors for training (RoboTool, GrScreenTool for EndoVis2017VOS experiments; FBMS, SegTrackV2 for DAVIS2016 experiments), with the state-of-the-art CIS approach (without and with post-processing, in parenthesis, taken from [YLL+21]) and a fully-supervised baseline ($\text{Baseline}_{\text{FS}}$). Mean IoU [%] and standard deviation are reported. The percentage of annotated training samples required by each method is also reported (Annot. [%]). Note that CIS (*) uses frames and optical flow to make predictions, while our approach only uses optical flow.

the results of surgical tool segmentation of individual frames. In order to evaluate model performance, mean Intersection-over-Union (IoU) between predictions and manually annotated ground truth (GT) is used.

3.4.1 Optical-Flow Segmentation

Optical flow segmentation by the *Teacher* network was evaluated on EndoVis2017VOS and DAVIS2016, and compared with a state-of-the-art deep learning approach for unsupervised Video Object Segmentation, called Contextual Information Separation (CIS, [YLSS19a]), adopting the same evaluation protocol on DAVIS2016 and providing freely available code to train it and test it in the surgical scenario. We report the CIS results both with and without post-processing, for fair comparison with our approach which does not make use of it, using the trained network parameters provided by the authors for DAVIS2016 experiments. Despite being trained using the PWC-net optical flow estimator [SYLK18], we observed that the CIS model provided more accurate results using RAFT-generated optical flow images: we thus reported results using the latter. On EndoVis2017VOS, the CIS model was trained from scratch, using the RAFT optical flow estimator: training was carried out using the code publicly released by the authors [YLSS19b]. We trained our *Teacher* model using RoboTool and GrScreenTool *shape-priors* for EndoVis2017VOS experiment, and SegTrackV2 and FBMS for DAVIS2016 experiment. We also report results of a fully-supervised baseline ($\text{Baseline}_{\text{FS}}$) model, having the same architecture as the *Teacher* network, trained on GT labels.

Experimental results, presented in Table 3.1, show that the proposed approach outperforms the state-of-the-art CIS approach (without post-processing) both in the surgical scenario (EndoVis2017VOS dataset) and in general object segmentation (DAVIS2016 dataset). The reason behind the significant improvement on EndoVis2017VOS (+16.32% ΔIoU) may reside in the independence of the proposed approach from strong a priori hypothesis on surgical tool and background motion (e.g. of incoherent background motion). In fact, our method lets the generator and discriminator adapt to the complexity of the optical flow domain, generating samples with possible cluttered background and partial tool occlusion, while still enforcing correct segmentation through the cycle-

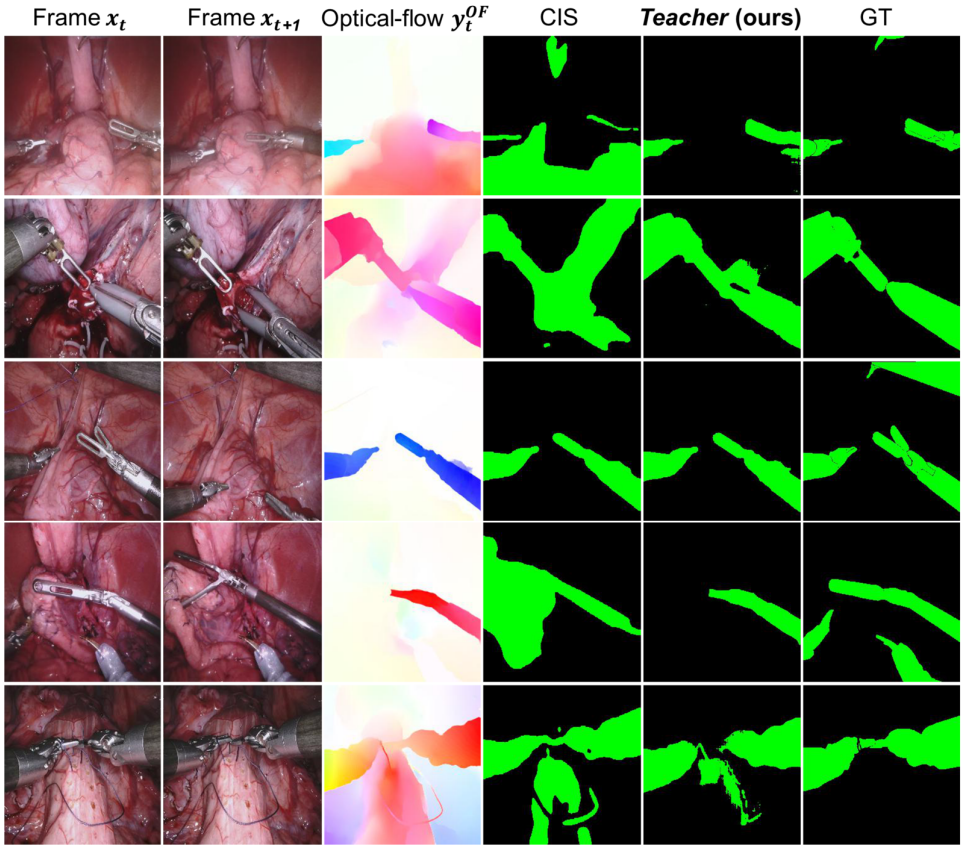

Figure 3.10: Optical flow segmentation on EndoVis2017VOS. Qualitative results showing frame couples used for optical flow computation, optical flow images after HSV standard conversion, predictions from CIS [YLSS19a] and Teacher (trained using RoboTool shape-priors), and ground truth (GT).

consistency loss. The only implicit constraint applied during training comes from the use of the *shape-priors*, which guides the *Teacher* towards segmenting *tool-shaped* regions. Since each optical flow image is normalized to have a maximum displacement of norm equal to 1, absolute tool motion does not directly impact the segmentation result. As a result, the *Teacher* model learns to segment *tool-shaped* regions characterized by a coherent motion, relatively different from the surroundings. This includes all those cases in which the tool is held still but the anatomic background moves due to physiological movements and nearby tool-tissue interactions. Examples of challenging generated optical flow images can be seen in Figure 3.16. Deeper insights on optical flow generation will be provided by the ablation study in Section 3.5.1. As a result, the optical flow segmenter becomes more robust to cluttered scenes, where tissue, as well as tools, moves coherently. As shown in the qualitative results from Figure 3.10, the proposed *Teacher* model outperforms the CIS approach especially when tools interact with the anatomy (e.g. pulling tissue, second row from bottom), allowing to drastically reduce the amount of background regions wrongly segmented as *tools*.

3.4.2 Single-frame Binary Tool segmentation

Single-frame binary tool segmentation was evaluated on the EndoVis2017Challenge, EndoVis2017VOS and STRAS datasets, according to the modalities reported in Section 3.3.1, using RoboTool and STRASMasks as *shape-priors*, respectively. For each experiment, we report results for the following networks:

- *Teacher* network, producing the pseudo-labels y_t^T from optical flow segmentation, evaluated against GT masks;
- *Proxy* network, directly trained on the noisy pseudo-labels, producing segmentation masks y_t^P from individual frames, evaluated against GT masks;
- *Student* network trained using *local* IoU masking, producing segmentation masks y_t^S from individual frames, evaluated against GT masks. The *Student* network is the output model of the proposed FUN-SIS approach.

For the EndoVis2017Challenge and EndoVis2017VOS experiments we compare the proposed approach with the unsupervised Anchor Generation and Semantic Diffusion (AGSD) approach [LWJ+20a], based on handcrafted features, and with the fully-supervised state-of-the-art approaches TernausNet-16 [SRKI18a] and MF-TapNet [JCDH19a]. Results on EndoVis2017VOS for these approaches were obtained by training the models using the code publicly released by the authors [LWJ+20b, SRKI18b, JCDH19b]. Additionally, we compare our results with $\text{Baseline}_{\text{FS}}$, a model sharing the same architecture as the *Student* network (Unet16), but trained in a fully-supervised way on the GT labels. We do not provide fully-supervised results on the STRAS dataset, due to the lack of GT training labels. We also do not provide results for the unsupervised AGSD approach, due to the fact that the handcrafted cues selected by the authors are specifically tailored for the EndoVis dataset, yielding poor results on the significantly different STRAS dataset. In addition to the standard IoU metric we also evaluate our solution using a Boundary-IoU, providing a better understanding of model behaviour on challenging boundary pixels.

Experimental results, reported in Table 3.2, show that the proposed approach enables to effectively train the *Student* network in a fully-unsupervised way, reaching 83.77% IoU on the EndoVis2017VOS dataset, 12.30% above the unsupervised AGSD approach and only 5.22% below the fully-supervised baseline. As hypothesized, the noise affecting the pseudo-labels generated by optical flow segmentation cannot be predicted from the individual frames, thus cannot be learnt by the *Proxy* network, which learns instead the *easiest* pattern compatible with the pseudo-labels, i.e. separating tools from anatomy. This results in a significant improvement of the *Proxy* network's predictions compared to pseudo-labels used for its training (+34.70% ΔIoU on EndoVis2017VOS). On top of this, the *Student* network significantly improves the segmentation quality, by training only on the *probably* well-labelled regions of the pseudo-labels, selected by means of the *local* IoU between pseudo-labels and *Proxy* predictions: the improvement of the *Student* network, with respect to the *Proxy* network, amounts to +8.99% ΔIoU on EndoVis2017VOS. Qualitative results presented in Figure 3.11 clearly show the dramatic improvement of the *Proxy* network compared to the *Teacher* network, and the refining effect of the *Student* network, producing accurate and sharp segmentation masks. In

	Annot. [%]	EndoVis2017VOS	EndoVis2017Challenge
TernausNet-16	100	(89.06±13.17)	83.60±15.83 (82.95±14.37)
MF-TAPNet	100*	(**89.61±13.22**)	**87.56±16.24** (85.81±15.94)
$\text{Baseline}_{\text{FS}}$	100	88.99±11.34	82.55±14.51
AGSD	0	(71.47±16.68)	67.85 (65.30±19.34)
Teacher (ours)	0	40.08±26.70	37.03±25.83
Proxy (ours)	0	74.78±14.99	68.31±19.11
Student (ours)	0	**83.77±12.28**	**76.25**±18.61

Table 3.2: Surgical tool segmentation of individual frames. Comparison of the proposed unsupervised method (trained using RoboTool shape-priors), with state-of-the-art unsupervised AGSD [LWJ+20a] approach, fully-supervised approaches TernausNet-16 [SRKI18a] and MF-TAPNet [JCDH19a], and fully-supervised baseline (Baseline_{FS}) on the EndoVis2017VOS and EndoVis2017Challenge datasets. Results in parenthesis for state-of-the-art approaches were obtained by training the models using the code released by the authors. Mean IoU [%] and standard deviation are reported. The percentage of annotated training samples required by each method is also reported (Annot. [%]). Note that MF-TAPNet uses 2 consecutive frames at inference time to make a prediction, while the other approaches use individual frames.

	p-value (t-test)	*Cohen's d*
Proxy-Teacher	$p << 0.001$	1.566
Student-Proxy	$p << 0.001$	0.612
$\text{Baseline}_{\text{FS}}$-Student	$p << 0.001$	0.448

Table 3.3: Statistical analysis of tool segmentation results obtained in EndoVis2017VOS (Table 3.2). For each pair, t-test was run (p-values reported in first column) and Cohen's d number was computed.

order to assess the statistical significance of the results on the EndoVis2017VOS dataset, pairwise t-tests were run (sample size N=1800) between *Proxy* & *Teacher*, *Student* & *Proxy* and $baseline_{FS}$ & *Student*, all showing statistically significant differences ($p << 0.001$ for all the three pairs). In addition, *Cohen's d* number was computed for such pairs, in order to quantify the strength of such statistically significant difference. *Cohen's d* numbers analysis, reported in Table 3.3, shows that the effect-size of such differences is *very large* between *Proxy* & *Teacher* ($d > 1.2$, $d = 1.566$), *medium/high* between Student & Proxy ($0.5 < d < 0.8$, $d = 0.612$) and *medium/small* between *fully-supervised baseline* & *Student* ($0.2 < d < 0.5$, $d = 0.448$) (according to [Coh13, Saw09]). In order to better understand the method's limitations we also analyzed the obtained segmentation results using a novel metric introduced in [CGD+21], called Boundary Intersection-over-Union (B-IoU). Compared to standard IoU, the most commonly used metric to evaluate segmentation results, B-IoU values only pixels close to the object boundaries, and is, as a result, more sensitive to boundary quality especially in larger objects. B-IoU first identifies the set of the input masks' pixels (prediction and GT masks) that are within a threshold distance d from each contour, and then computes the IoU of these two sets. Threshold d determines the relative weight of the contour (a smaller d value weights the boundaries more). We set it to 2% of the mask diagonal, as in [CGD+21]. Qualitative results breaking down B-IoU computation and quantitative results comparing FUN-SIS with the unsupervised AGSD method and the fully-supervised baseline are shown in Figure 3.12 and Table 3.4,

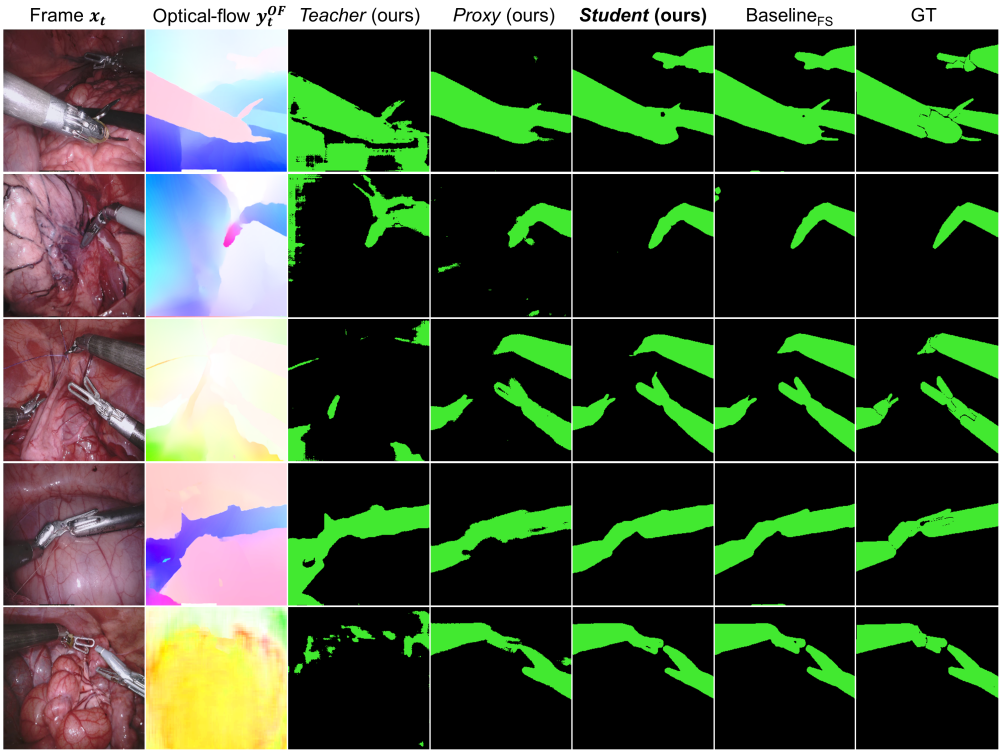

Figure 3.11: Surgical tool segmentation on the EndoVis2017VOS dataset. Qualitative results showing, from left to right, input frame I_t, optical flow image y_t^{OF} using HSV standard conversion, predictions from Teacher (using RoboTool shape-priors), Proxy, Student and fully-supervised baseline (Baseline$_{FS}$), and ground truth (GT).

respectively.

	Annot. [%]	B-IoU
Baseline$_{FS}$	100	76.38±11.43
AGSD	0	50.11±14.48
Teacher (ours)	0	31.56±19.69
Proxy (ours)	0	55.73±15.32
Student (ours)	0	68.46±12.62

Table 3.4: Surgical tool segmentation of individual frames. Results of the proposed method on the EndoVis2017VOS dataset using RoboTool shape-priors. Mean Boundary-IoU (B-IoU) [%] and standard deviation are reported. The percentage of annotated training samples required by each method is also reported (Annot. [%]).

B-IoU results follow the same trend as standard IoU results, but help highlight the gap between *Student* and AGSD (+18.35% ΔB-IoU) and between *Student* and fully-supervised baseline (-7.92% ΔB-IoU). We believe that this gap is mainly due to the sub-optimality of the optical flow estimation model to capture fine-scale details (Figure 3.13). Indeed, while the *Proxy* network's segmentation masks can be imprecise at object boundaries due to the way the network is trained, they do not directly supervise the *Student* network training: instead, they are used to select *probably* well-labelled regions of the pseudo-labels, where the *Student* network gets actually trained. Therefore, the quality of the

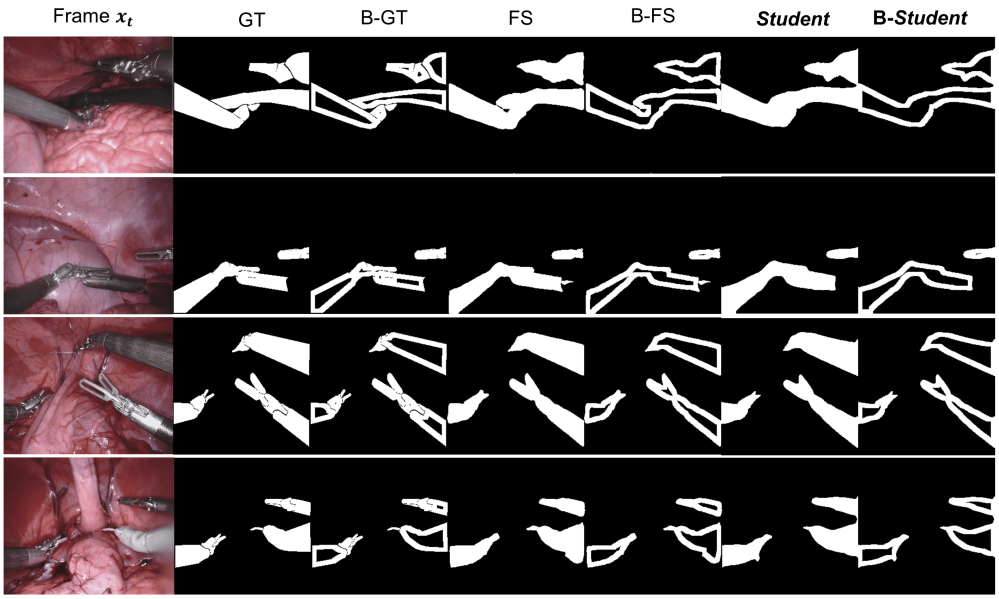

Figure 3.12: Surgical tool segmentation on the EndoVis2017VOS dataset: example of boundary masks used for Boundary IoU computation. From left to right, input frame I_t, ground truth mask (GT) and boundary version (B-GT), fully-supervised baseline mask (FS) and boundary version (B-FS), Student mask (Student) and boundary version (B-Student). Boundary IoU is computed as a standard IoU between B-GT and predicted boundary mask (e.g. B-Student).

pseudo-labels at fine scale directly impacts the *Student*'s performance. Potential improvements are discussed in Section 3.6.

As expected, the performance on the EndoVis2017Challenge dataset, where tools such as the Ultra-Sound probe are considered as part of the *background* class, is lower than the one on EndoVis2017VOS, while still outperforming the unsupervised AGSD approach (+8.40% ΔIoU). This is due to the fact that our approach, despite not being trained using specific *shape-priors* of these tools, is still able to generalize and segment them together with the da Vinci ones. Examples of frames containing the drop-in Ultra-Sound probe are shown in Figure 3.11, first and fourth row from the top. In order for our approach to learn such semantic discrimination between the two *instrument* classes, pure motion information may not be sufficient. The possible extension of FUN-SIS to multi-class segmentation will be discussed in Section 3.6. We also analyze the difference between the 2-step and 3-step training strategies described in Section 3.2.4. Results, shown in Figure 3.14, confirm that the two modalities provide comparable results, as suggested in Section 3.2.4. We thus consider the 2-step approach superior, due to the shorter training time required. Results obtained on the challenging STRAS dataset, reported in Table 3.5, confirm the ability of the method to effectively learn surgical tool segmentation in a fully-unsupervised way. The *Student* network, trained without any domain-specific hyper-parameter tuning, reaches an IoU equal to 66.37%, despite being trained on very low-quality pseudo-labels (29.93% IoU). As observable from Figure 3.15, in fact, optical flow images appear less sharp compared to the EndoVis2017 ones, mainly due to image blurring and lower image resolution, influencing the overall performance. The implications of the method's dependency on optical flow quality will be discussed in Section 3.6. Additional qualitative results for the *Student* network on the EndoVis2017VOS and STRAS datasets are displayed in Figures 3.28&3.29 at the end of

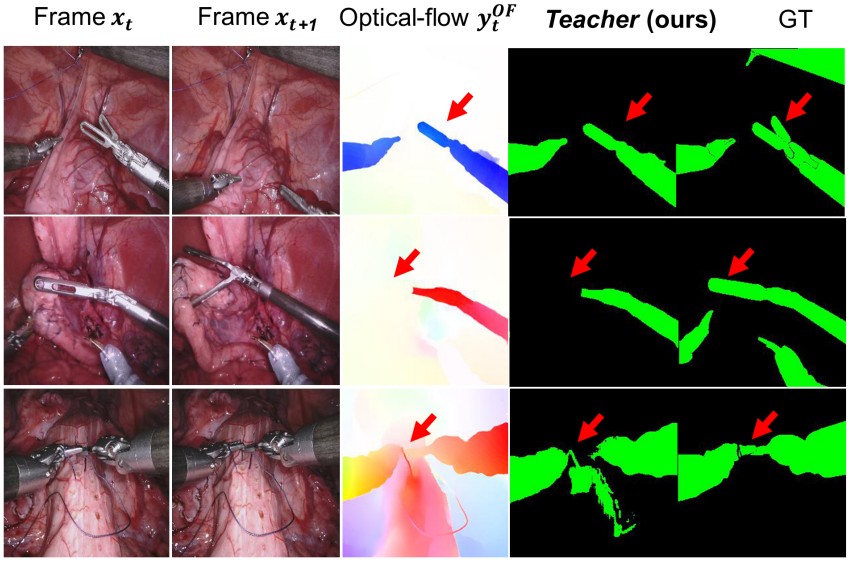

Figure 3.13: Optical flow estimation inaccuracy leading to missing details in the pseudo-labels (Teacher network predictions) such as tool tips.

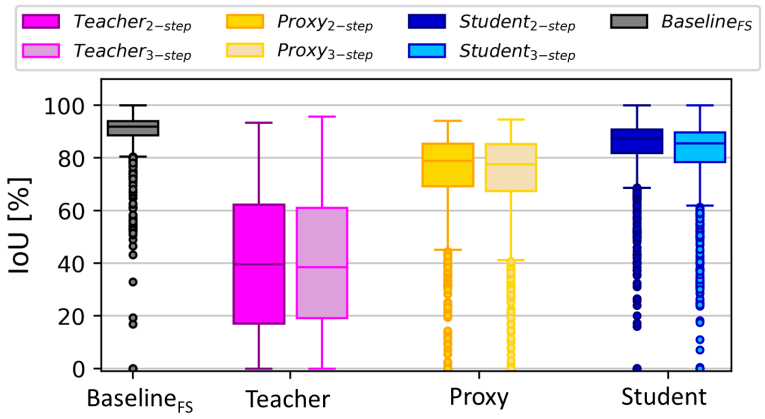

Figure 3.14: Box-plots showing IoU distributions from EndoVis2017VOS segmentation experiment (Table 3.2). Fully-supervised baseline Baseline$_{FS}$ (grey), Teacher (purple 2-step, light purple 3-step), Proxy (yellow 2-step, light yellow 3-step), Student (blue 2-step, light blue 3-step).

the Chapter.

	Annot. [%]	STRAS
Teacher (ours)	0	29.93±8.51
Proxy (ours)	0	55.07±6.47
Student (ours)	0	**66.37±5.14**

Table 3.5: Surgical tool segmentation of individual frames. Results of the proposed method on the STRAS dataset using STRASMasks shape-priors. Mean IoU [%] and standard deviation are reported. The percentage of annotated training samples required by each method is also reported (Annot. [%]).

3.5 Ablation Studies and Additional Experiments

In order to provide a more in-depth understanding of the proposed FUN-SIS approach, we performed several ablation studies on crucial aspects of the method. In the next paragraphs, each experiment is followed by a short discussion of the obtained results, in order to facilitate the reading.

3.5.1 Optical-Flow Augmentation and Noise Vector Size

We first analyze optical flow surgical tool segmentation by the *Teacher* network. In particular, we evaluate the impact of the two proposed strategies to tackle the *complexity-imbalance* between optical flow and *shape-priors* domain in the generative part of the *Teacher* training, described in Section 3.2.1: noise concatenation and optical flow augmentation $AugmFlow$. We trained the *Teacher* model using different sizes of the concatenated noise vector $\mathbf{n}$, with and without the optical flow augmentation $AugmFlow$.

Qualitative and quantitative results are shown in Figures 3.16 and 3.17, respectively. Quantitaive results highlight how optical flow augmentation $AugmFlow$ plays a crucial role in counteracting *complexity-imbalance*, allowing to reach quasi-optimal performance even without noise concatenation (continuous line, “no-noise"). Noise concatenation also appears effective, with peak *Teacher* performance reached with noise size 32 and $AugmFlow$. We did not notice any significant improvement with larger noise vector sizes. From qualitative results shown in Figure 3.16, it can be noted how noise concatenation allows to both generate more realistic and variable optical flow images and disentangle tool configurations and optical flow appearance. Note how, when changing *shape-priors*, optical flow image appearance changes when noise is not concatenated (x0, first block), but remains similar in case of noise concatenation (x1 and x32, second and third block, respectively). It can also be observed how the most variable results are obtained with a noise vector size of 32 (third block, x32), with complexity increasing from leftmost column (noise vector of zeros, more frequently sampled during training from the normal distribution) to rightmost column (noise vector of ones, rarely sampled during training), where tools are hardly recognizable and background appears to feature more consistent motion.

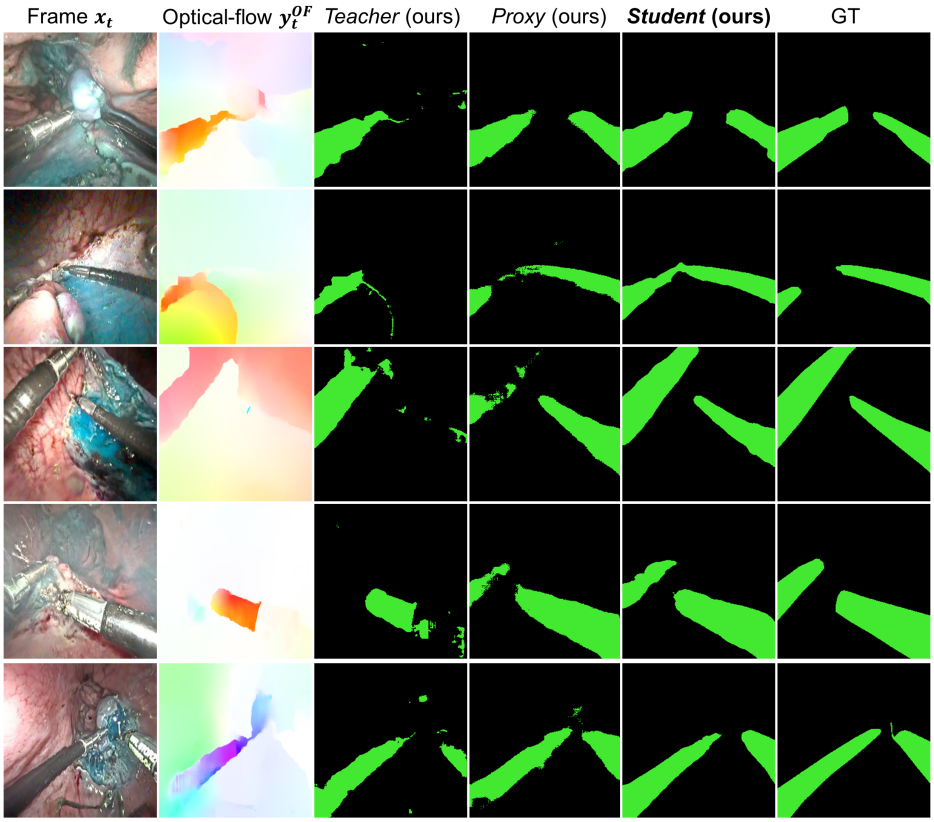

Figure 3.15: Surgical tool segmentation on the STRAS dataset. Qualitative results showing, from left to right, input frame I_t, optical flow image y_t^{OF} using HSV standard conversion, predictions from Teacher (using STRASMasks shape-priors), Proxy and Student, and ground truth (GT).

3.5.2 Proxy Network Architecture

In Section 3.2.2 we hypothesized the benefit of a limited *Proxy* network capacity, in order to encourage the learning of the *easiest* pattern shared between training samples, compatible with the pseudo-labels. We investigate this hypothesis by evaluating the performance of *Proxy* and *Student* networks, when using different *Proxy* architectures (Unet11 and Unet16, defined in Section 3.3.3) on the EndoVis2017VOS dataset.

Results shown in Figure 3.18 confirm that a shallower *Proxy* network (Unet11) learns more effectively from the pseudo-labels than a deeper one (Unet16), quantified in an improvement of +5.44% ΔIoU. Additionally, this study provides the experimental proof that the *Student*'s improvements with respect to the *Proxy* are not due to their different architectures. Indeed, when using the same architecture for both of them (Figure 3.18, left) the *Student* still outperforms the *Proxy* by a large margin (+10.61% ΔIoU).

3.5.3 Loss Function Coefficients (α_P, α_S)

We investigate the impact of the balancing factors α_P and α_S between cross-entropy (CE) and log IoU losses in *Proxy* and *Student* networks training (Equations 3.8&3.13). In our experiments we consider the case $\alpha_P = \alpha_S = \alpha$, with α ranging from 0 (only CE loss) to 1 (only log IoU loss).

Results shown in Figure 3.19 highlight the positive impact of log IoU loss, especially

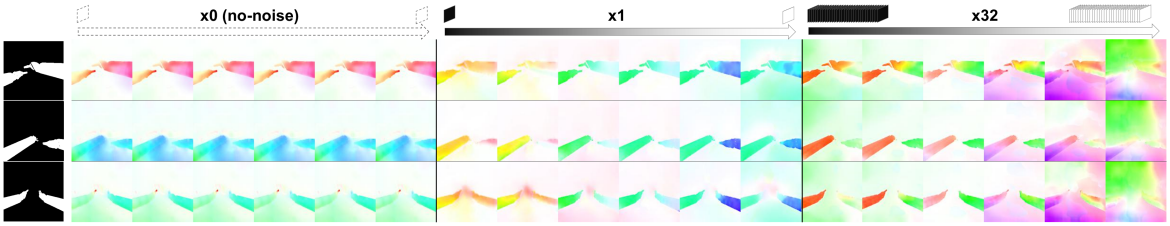

Figure 3.16: Qualitative results of the optical flow generator (G), trained using different size of input noise vector among {no-noise,1,32}. First column: input shape-priors; first block (x0), no noise concatenation; second block (x1), noise vector of size 1; third block (x32), noise vector of size 32. For each of the 3 blocks, from left to right, the noise vector was smoothly interpolated between all zeros to all ones (trivial for x0, having no concantenated noise).

.

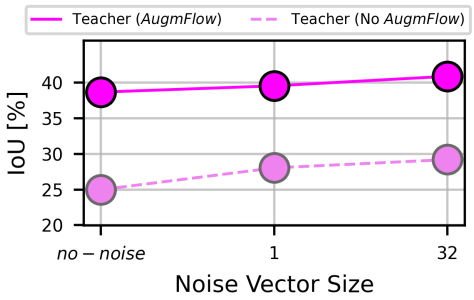

Figure 3.17: Analysis of the impact of noise vector size (no-noise, 1, 32) and flow augmentation $AugmFlow$ on optical flow segmentation results by the Teacher network on EndoVis2017VOS. Mean IoU [%] is reported.

on the *Proxy* network (+19.79% ΔIoU improvement between $\alpha = 1$ and $\alpha = 0$). This can be in part explained by the diminished sensitivity of IoU-based losses due to class-imbalance. However, the greater improvement brought by the log IoU loss to the *Proxy* network, directly trained on raw pseudo-labels, compared to the *Student* network, may suggest that the log IoU loss is more robust to the noise of motion-derived pseudo-labels. Additional in-depth studies are required to investigate this hypothesis.

3.5.4 Local IoU Parameters' Impact

While state-of-the-art *learning-from-noisy-labels* approaches usually require a *Teacher* model trained on clean labels in order to identify well-labelled regions of noisy pseudo-labels, we perform this search in a fully-unsupervised way. As detailed in Section 3.2.3, *probably* well-labelled regions are selected according to the *agreement* between the pseudo-labels (*Teacher* model's predictions from optical flow segmentation y_t^T) and *Proxy* model's predictions y_t^P. The agreement is measured by the *local* IoU, parametrized by the window size w ($w = h$ in our experiments), and binarized through the threshold parameter ϵ_{IoU}, representing the minimum agreement required to consider a region well-labelled. The choice of these two parameters influences 1) the *effective* number of pixels on which the *Student* network is trained, 2) the average *effective* IoU (IoU_{eff}) of the training labels, defined as the IoU between ground truth masks GT and pseudo-

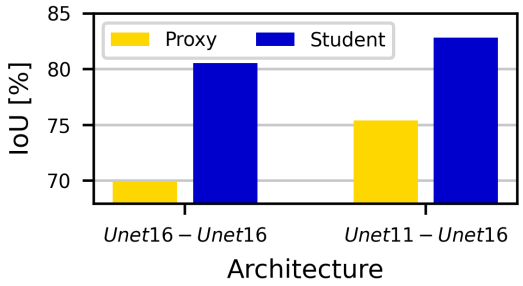

Figure 3.18: Analysis of the impact of Proxy network's architecture on surgical tool segmentation results of Proxy (yellow) and Student (blue) networks, on EndoVis2017VOS. Mean IoU [%] is reported.

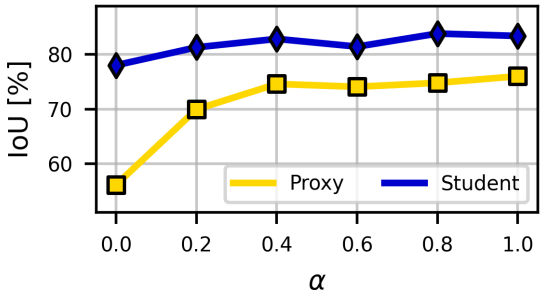

Figure 3.19: Analysis of the impact of loss function balancing coefficients α_P and α_S on Proxy (yellow) and Student (blue) networks, on EndoVis2017VOS. We only consider the case $\alpha_P = \alpha_S = \alpha$; α equal 0 corresponds to cross-entropy loss only, α equal 1 corresponds to log IoU loss only. Mean IoU [%] is reported.

labels y_t^T, computed only for the selected regions according to the binarized *local* IoU ($\overline{IoU}^{loc}_{(w,h)}$) between y_t^T and y_t^P:

$$IoU_{eff} = \frac{|(GT \cap y_t^T) \cap \overline{IoU}^{loc}_{(w,h)}|}{|(GT \cup y_t^T) \cap \overline{IoU}^{loc}_{(w,h)}|}. \tag{3.14}$$

We evaluate the influence of w and ϵ_{IoU} on the *effective* training size (expressed as total number of selected pixels over total number of pixels in the training dataset) and on the average IoU_{eff} in the training dataset. For these experiments, we considered trained *Teacher* and *Proxy* models on EndoVis2017VOS. We then varied ϵ_{IoU} and w in a grid-like manner, with ϵ_{IoU} ranging from 0.0 to 1.0 with a step equal to 0.05, and w in $\{1,2,4,8,16,32,64,128,256\}$. For each couple (w,ϵ_{IoU}) we then evaluated *effective* training size and average IoU_{eff} on the EndoVis2017VOS training set, in order to provide an insight of the *effective* training carried out.
Experimental results shown in Figure 3.20 confirm that the agreement between pseudo-labels (optical flow segmentation masks from the *Teacher*) and *Proxy* predictions is directly correlated to the quality of the pseudo-labels. Figure 3.20 (right) shows the positive correlation between *Proxy-Teacher* agreement (ϵ_{IoU}) and average *effective* IoU, especially for large window sizes w of the local IoU operation. As expected, the

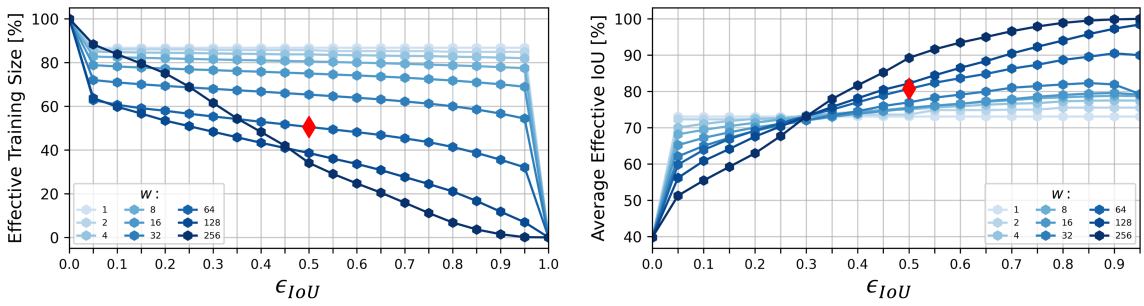

Figure 3.20: Impact of local IoU parameters (ϵ_{IoU} and window size w) on effective training size (left) and average effective IoU (right). x-axis can be interpreted as the level of agreement between Teacher and Proxy required in order to select a certain region (e.g. with ϵ_{IoU} equal to 0.8 a region is considered well-labelled only if the IoU between Proxy and Teacher predictions for that region is at least 80%). Red markers correspond to $w = 64$ and $\epsilon_{IoU} = 0.5$, the values used in our main experiments.

experiment also shows that requiring higher agreement reduces the amount of data effectively used for training, with a similar but inverse relationship. This creates the need to properly select the two parameters in order to train the *Student* network on good quality labels, in order to facilitate convergence, while keeping the training set large enough to allow generalization and robustness. In light of this experiment, the values of window size w and ϵ_{IoU} chosen for experimental validation, respectively 64 and 0.5, represent a good compromise, allowing to train the *Student* network on 50.48% of the total training data on EndoVis2017VOS, with an effective IoU of the pseudo-labels equal to 80.70% (high-quality labels).

3.5.5 Shape-Priors Quality & Quantity

Shape-priors represent the only external information required by the proposed approach for training. In order to investigate their impact on the whole training process, we performed two sets of experiments. First, we evaluated the performance of our models (*Teacher*, *Proxy*, *Student*) when trained using RoboTool and GrScreenTool *shape-priors*, on the EndoVis2017VOS dataset, in order to evaluate the impact of different sources (i.e. *recycled* annotations from a different dataset and automatically segmented tools from green-screen recordings); secondly, we trained our models using different percentages of the available RoboTool *shape-priors*, from 100% to 1%, with and without on-the-fly augmentation $AugmMask$.

Shape-Priors	RoboTool	GrScreenTool
Teacher (ours)	40.08±26.70	40.47±25.62
Proxy (ours)	74.78±14.99	73.63±15.11
Student (ours)	83.77±12.28	82.63±13.01

Table 3.6: Analysis of the impact of the shape-priors dataset on frame segmentation. Comparison of the proposed method trained using RoboTool and GrScreenTool as shape-priors on EndoVis2017VOS. Mean IoU [%] and standard deviation are reported.

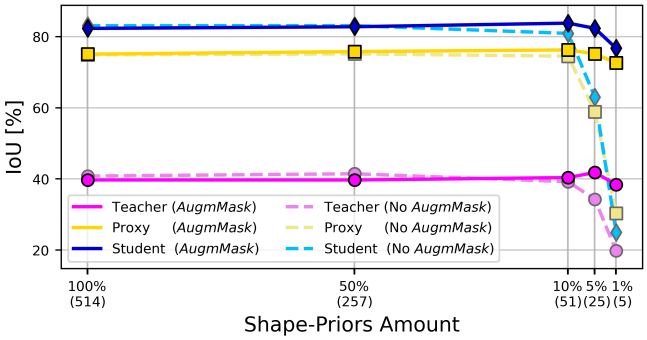

Figure 3.21: Analysis of the impact of decreasing shape-priors quantity on individual frame and optical flow segmentation, with and without $AugmMask$ augmentation, on EndoVis2017VOS. On the x-axis, the amount of RoboTool shape-priors used for training is reported (absolute number and percentage with respect to the total number). Mean IoU [%] for Student (blue; dashed: trained without $AugmMask$), Proxy (yellow; dashed: trained without $AugmMask$), Teacher (purple; dashed: trained without $AugmMask$) is reported.

Experimental results highlight how our FUN-SIS approach is extremely robust to varying quantity and source of the *shape-priors*. Experiments using GrScreenTool, reported in Table 3.6, provide comparable performance to the ones using RoboTool, despite the significantly different appearance of tools, as shown in Figure 3.8. In addition, experiments on *shape-priors* quantity (Figure 3.21), show how the performance of *Teacher*, *Proxy* and *Student* remains optimal even when using as few as 51 RoboTool *shape-priors* masks (10% of total) for training. If augmented on-the-fly using the $AugmMask$ protocol (random cropping and flipping), RoboTool *shape-priors* can be further reduced to a total number of 5 instances (1% of total), with limited performance drop (-5.57% ΔIoU compared to 100% case).

3.5.6 Noise Properties (Unpredictability & Polarization)

We investigate the impact of the *unpredictability* and *polarization* properties presented in Sections 3.2.2 and 3.2.3 on the proposed *learning-from-noisy-labels* approach. To this aim, we carried out experiments with artificially controlled type and intensity of noise affecting the pseudo-labels, as described in Section 3.3.2. We then substituted the pseudo-labels y_t^T, in our training pipeline, with the corrupted EndoVis2017VOS labels and trained the *Proxy* and the *Teacher* networks according to the same modalities as the previous experiments. The three noise strategies presented in Section 3.3.2 were designed to highlight the effect of the *unpredictability* and *polarization* properties. In *Systematic-Erosion* experiment, each mask was eroded, making the noise signal *predictable* and *not-polarized* (all tools are equally affected by the noise); in *Erosion&Dilation* experiment, each mask was either randomly eroded or dilated, making the noise signal *unpredictable*, but still *not-polarized* (each tool mask is affected by an error, either due to erosion or dilation); finally, in *Tool-Drop* experiment, individual tools were either perfectly annotated or not annotated at all, making the noise signal both *unpredictable* and *polarized*. Differently from the real scenario of pseudo-labels derived from optical flow segmentation, no false-positives are present in the *Tool-Drop* experiment, and the *polarization* is perfect (tools are either fully present or completely

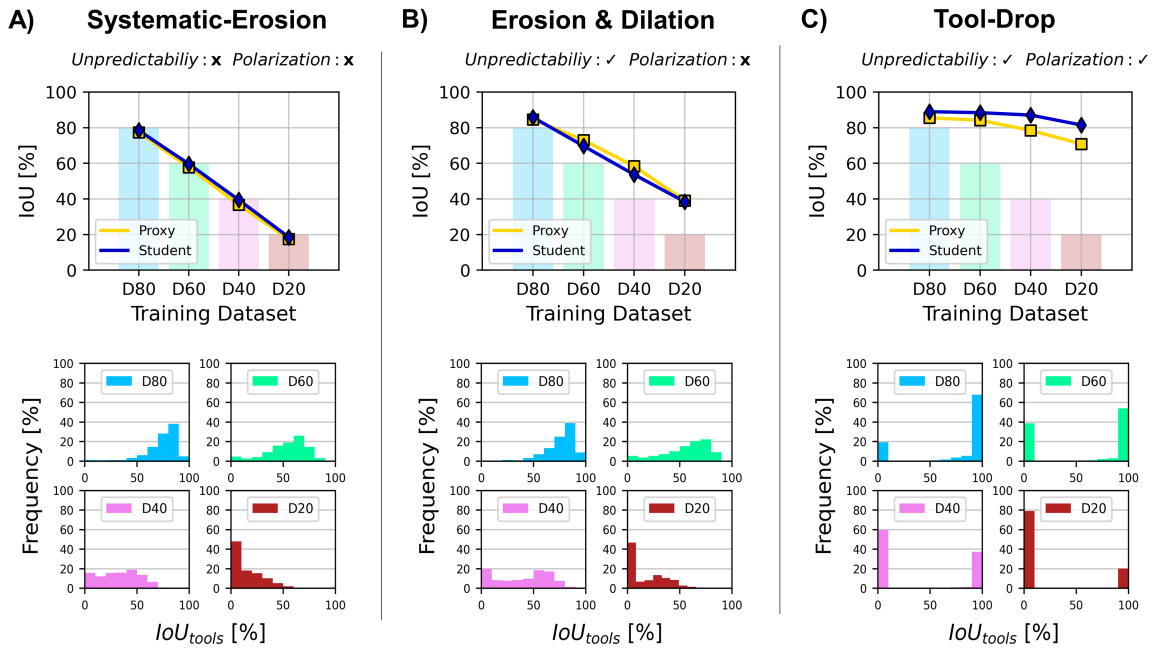

Figure 3.22: Analysis of the impact of unpredictability and polarization noise properties on the proposed method, on the artificially-corrupted EndoVis2017VOS datasets. Top: for each of the 3 noise sources (A, Systematic-Erosion, predictable and not-polarized; B, random Erosion & Dilation, unpredictable and not-polarized; C Tool-Drop, unpredictable and polarized) Proxy (yellow) and Student (blue) models were trained on the EndoVis2017VOS training dataset, having ground truth labels corrupted with different levels of such noise. The colored bars are meant to improve readability, by visually showing the mean IoU between each training dataset labels and ground truth clean labels (~80% for D80, ~60% for D60, ~40% for D40, ~20% for D20); Bottom: for each set of noisy labels, per-tool IoU histograms (IoU_{tools}) computed as shown in Figure 3.23, are reported.

dropped), while in the real case tools can also be partially segmented.

Results of the conducted experiments (Figure 3.22) clearly highlight the impact of the two noise properties, as well as the ability of the proposed solution to leverage them. When the noise is predictable (Figure 3.22-A, top), the *Proxy* network can perfectly learn to fit it, even when the corruption is minimal (D80). Contrarily, when noise cannot be inferred from single frames (Figure 3.22-B&C, top), the *Proxy* network, unable to learn the noise pattern, will learn the *easiest* general pattern compatible with the labels, resulting in significantly better predictions than the noisy labels used for its training (on average, +13.76% ΔIoU in *Erosion&Dilation*, +29.75% ΔIoU in *Tool-Drop*). The effectiveness of the *Student* network training is instead mainly influenced by the *polarization* property. When the noise is not polarized (Figure 3.22-A&B, top), the *Student* network does not benefit from region selection through *local* IoU (+1.69% and -1.87% ΔIoU, respectively, of *Student* compared to *Proxy* network). Instead, when the noise is polarized, well-labelled regions can be effectively identified using *local* IoU, allowing for a consistent improvement of *Student* predictions, compared to *Proxy* ones (+6.73% ΔIoU on average, +8.60% ΔIoU in D40). The improvement is aligned with the one obtained in the experiments from Section 3.4.2 (+8.99% ΔIoU), where the pseudo-labels were produced via unsupervised surgical tool segmentation by the *Teacher* network and had an IoU with the GT equal to 40.08%. Overall, the proposed approach allows to maintain an IoU of at least 81.49% (compared to the 88.99% reached by fully-supervised training of the *Student* model on

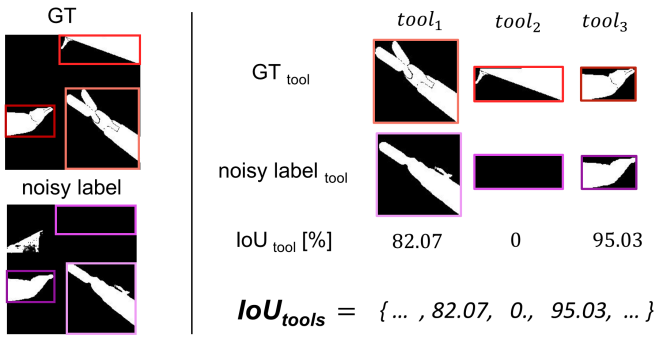

Figure 3.23: Computation of per-tool IoU between ground truth masks and noisy labels. Left: example of ground truth mask (GT) and noisy label. The smallest region containing each tool in the GT mask is extracted; the same exact region is extracted from the noisy label. Right: Intersection-over-Union (IoU_{tool}) is computed between each region extracted from GT (GT_{tool}) and noisy label ($noisy\ label_{tool}$); the process is repeated for each tool in each frame of the dataset, and each IoU_{tool} is stored in IoU_{tools}. The distribution of per-tool IoU can then be visualized through histogram plots (Figures 3.22&3.24).

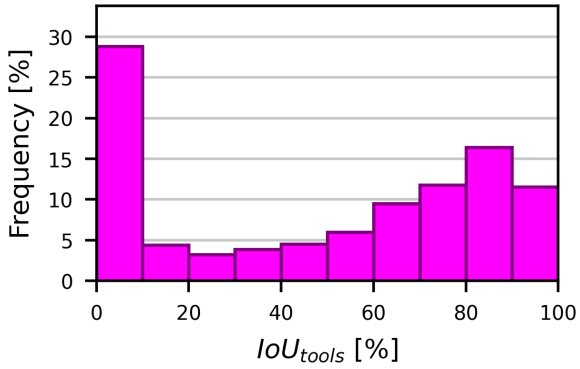

Figure 3.24: Per-tool IoU histogram (IoU_{tools}), computed as shown in Figure 3.23, for pseudo-labels derived from motion segmentation by the Teacher model on EndoVis2017VOS. Note how the distribution tends to be polarized on leftmost bin (completely mislabelled tools) and rightmost bins (almost-perfectly segmented tools).

clean labels, Table 3.2), even when trained on extremely low-quality training labels (Figure 3.22-C, top: *Tool-Drop*, D20 i.e. ~20% IoU between training labels and GT). When trained on D80 and D60, the *Student* network reaches optimal performance (88.98% and 88.41% IoU, respectively).

In order to provide a direct visualization of the *polarization* property, we also report, for each set of noisy labels, including the motion-derived pseudo-labels by the *Teacher* model, per-tool IoU histograms (IoU_{tools}). Per-tool IoU can be computed, as shown in Figure 3.23, by extracting the smallest regions containing each tool from the GT labels, and computing the IoU between this region and the corresponding one from the corresponding pseudo-label. This process, while approximate (an extracted region from GT label may contain more than one tool), allows to produce a clear visualization of the *polarization* property, by plotting the histogram of the obtained IoU_{tools}. Histograms are shown in Figure 3.24, for motion-derived pseudo-labels, and in Figure 3.22, bottom, for artificially corrupted labels. From Figure 3.22, bottom, it is possible to intuitively com-

	BF	PF	ND	VS	GR	S	O
Baseline_{FS}	78.68	84.29	90.39	89.88	86.39	92.42	87.74
Teacher	33.63	33.58	30.67	36.80	19.69	30.71	51.59
Student	73.57	78.66	83.93	85.91	82.07	86.19	83.01

Table 3.7: Per-tool segmentation results on the EndoVis2017VOS dataset separated by surgical tool class: Bipolar Forceps (BF), Prograsp Forceps (PF), Needle Driver (ND), Vessel Sealer (VS), Grasper Retractor (GR), Scissors (S), Other (O). Mean IoU [%] is reported for fully-supervised baseline (Baseline_{FS}), Teacher and Student networks.

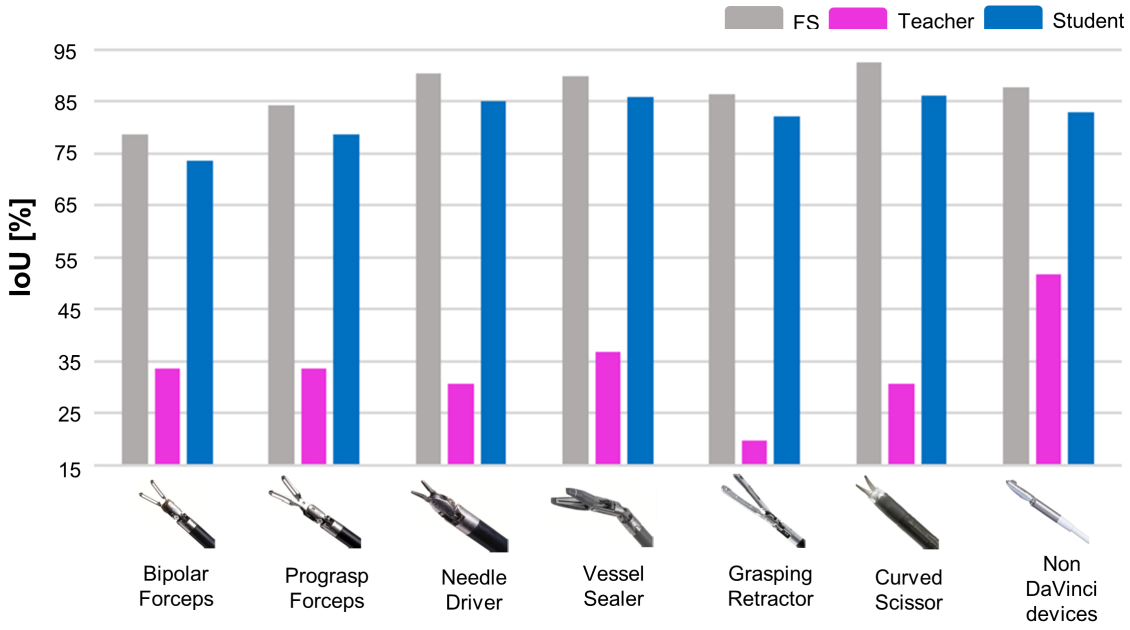

Figure 3.25: Break-down of the per-tool IoU across the 7 tool classes present in EndoVis2017, for fully-supervised baseline (FS), Teacher and Student networks.

pare the case of not-polarized noise (A,B), where $\text{IoU}_{\text{tools}}$ values are mostly distributed around a single peak, to polarized noise (C), where the values appear concentrated on leftmost bin (full tool annotations missed) and rightmost bin (perfectly labelled tools). In the case of pseudo-labels derived from optical flow segmentation (Figure 3.24), the histogram, despite being smoothed by the sub-optimality of optical flow estimator and segmenter described in Section 3.2.3, still displays the *polarization* property, allowing efficient *Student* network training.

3.5.7 Per-class IoU evaluation

Given the per-tool IoU metric defined above, segmentation results from the main experiment (Table 3.2) are here analyzed by breaking them down into the 7 different tool categories present in EndoVis2017 dataset.
Results, reported in Table 3.7 and visualized in Figure 3.25, highlight how, as expected, the quality of optical flow segmentation by the Teacher network decreases for passive tools, like the *grasping retractor*, often used to hold anatomical structures still. It is worth noticing that during training the *Teacher* network learns to segment tool-shaped regions

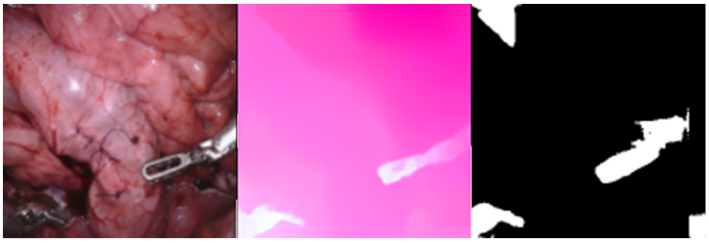

Figure 3.26: Surgical tool segmentation by the Teacher network in the case of static tools: physiological anatomical movements highlight surgical tools, even while being held still, allowing a proper segmentation by the Teacher network.

characterized by a coherent motion, relatively different from the surroundings. In principle, anatomical structures always feature some degree of motion (physiological movement, deformation from nearby interactions), inconsistent with the surgical tool motion (or non-motion), allowing the *Teacher* model to properly segment them (Figure 3.26). However, given a sub-optimal optical flow estimator, small physiological motions may be missed. This could be problematic for passive tools like graspers, as proved by the quantitative results in Table 3.7. Nonetheless, the proposed *learning-from-noisy-labels* strategy manages to handle this noise, allowing to properly train the *Student* network to achieve a final segmentation performance for these tools comparable to the other tools.

3.5.8 Random Unlabelled Data

In order to show the ease-of-use and robustness of the proposed FUN-SIS approach, we trained our models on the surgical robotic dataset RandSurg, described in Section 3.3.1 and tested on EndoVis2017VOS. The RandSurg dataset was created by collecting random public videos of surgical procedures, and performing minimal data curing. Training was carried out according to the same modalities as the other experiments, using RoboTool *shape-priors* and varying amounts of the RandSurg data, ranging from very few (31 i.e. 1% of total available) to all the available frames (3136).

Experimental results shown in Figure 3.27 show that, despite the limited data curing and pre-processing of the input data, the method can easily leverage the increasing amount of available data to effectively train the models. The *Student* network reaches a peak IoU equal to 79.65% on EndoVis2017VOS, comparable to the 83.77% obtained when training on unlabelled data from the same dataset (Table 3.2).

3.5.9 FUN-SIS Applicability on another Domain: Cholec80

We demonstrate the applicability of the proposed FUN-SIS approach on a different domain than the robotic one it was validated on. To this aim, we trained and qualitatively tested our *Student* model on the unlabelled Cholec80 dataset, consisting of manual laparoscopic cholecystectomy procedures. Training was carried out using RoboTool *shape-priors*, despite the different appearance of tools between robotic and manual laparoscopic videos.

Results shown in Figure 3.30 qualitatively confirm that the proposed method is applicable to a different surgical domain, even without domain-specific hyper-parameters

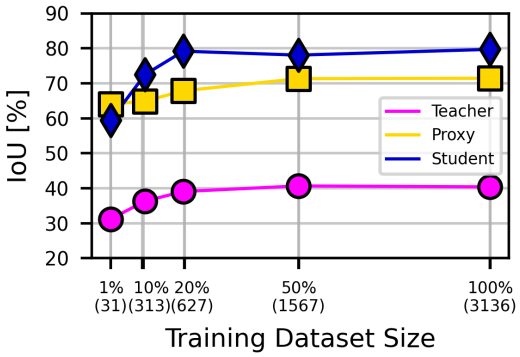

Figure 3.27: Analysis of proposed method performance when trained on increasing amounts of unlabelled RandSurg data, a dataset consisting of randomly selected surgical videos, downloaded from the public repository [Wor], and tested on EndoVis2017VOS. On the x-axis, the amount of RandSurg frames used for training is reported (absolute number and percentage with respect to the total number). Mean IoU [%] for Student (blue), Proxy (yellow), Teacher (purple) is reported.

tuning and with minimal pre-processing. Furthermore, they prove that despite the differences between *shape-priors* and target tools, segmentation can still be effectively carried out. In future work, we plan to quantitatively evaluate FUN-SIS generalization ability, especially for cases, like the *RoboTool*-Cholec80 one, where *shape-priors* and unlabelled videos come from significantly different domains. This will require building a benchmark dataset, manually annotated for surgical tool segmentation, containing data from different surgical procedures, performed with both manual and robotic-assisted laparoscopic techniques.

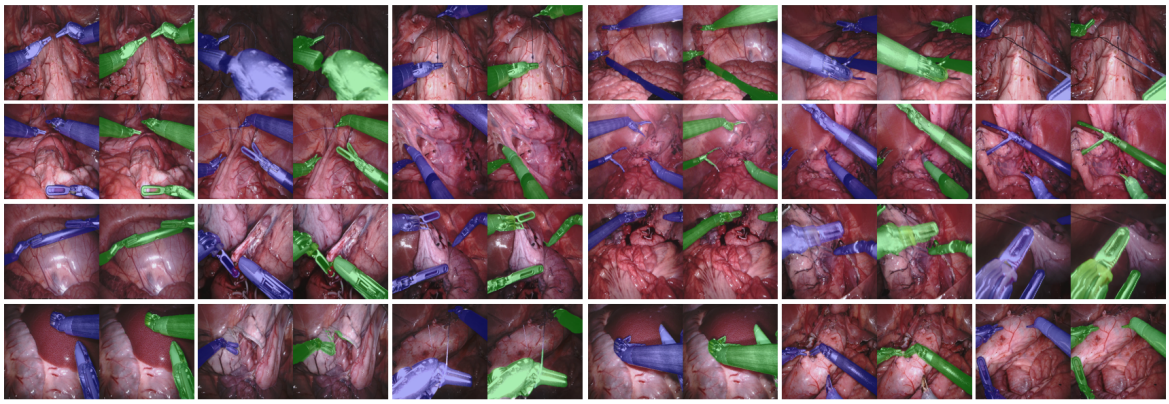

Figure 3.28: Qualitative results on the EndoVis2017VOS dataset, from the experiment reported in Table 3.2. Original frame overlapped with ground truth (blue) and Student network's prediction (green).

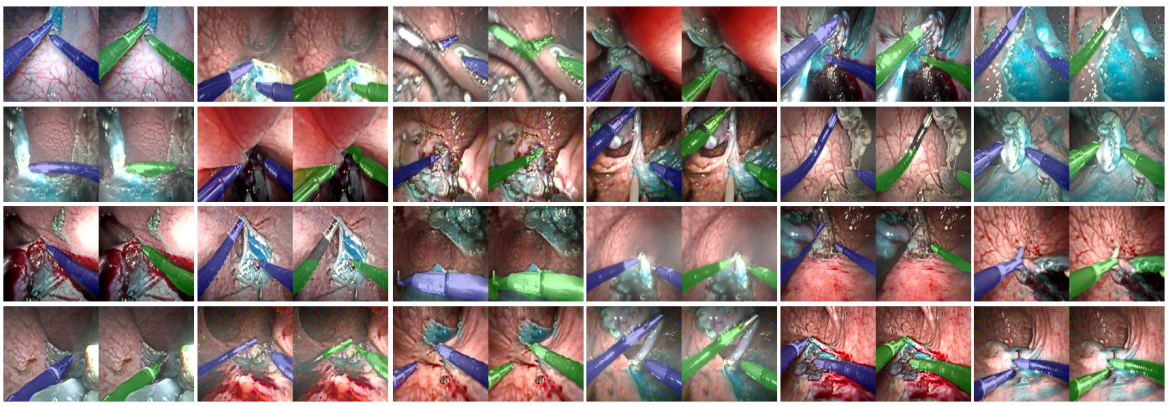

Figure 3.29: Qualitative results on the STRAS dataset, from the experiment reported in Table 3.5. Original frame overlapped with ground truth (blue) and Student network's prediction (green).

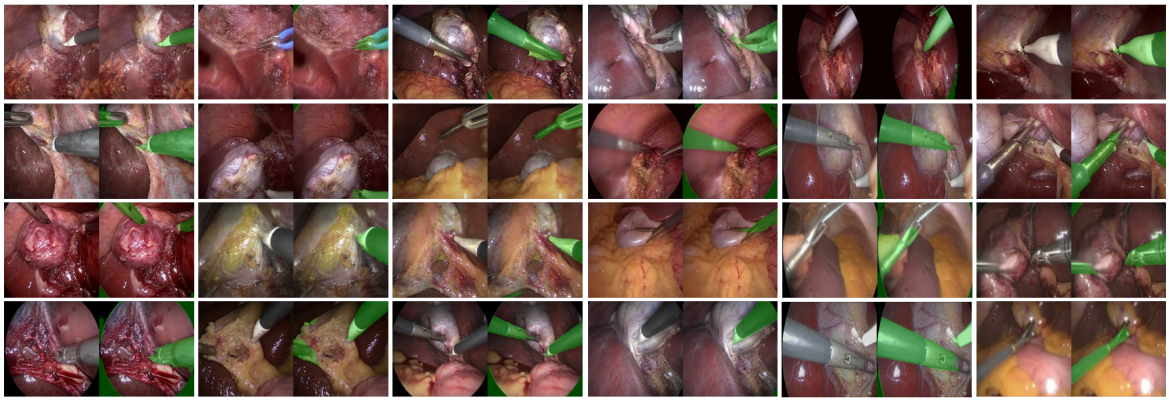

Figure 3.30: Qualitative results on the Cholec80 dataset. Original frame and overlapping between Student network prediction and original frame are shown. Training was carried out using RoboTool shape-priors.

3.6 Discussion and Future Work

In order to validate the proposed FUN-SIS approach, several experiments were performed and presented, including optical flow segmentation (Section 3.4.1), single-frame segmentation (Section 3.4.2, main experiment) and several ablation studies (Section 3.5), dissecting the method and highlighting its key aspects. The obtained results strongly support the soundness of FUN-SIS: binary surgical tool segmentation was effectively carried out in various datasets including EndoVis2017 (robotic surgery), STRAS (flexible endoscopic surgery), and Cholec80 (manual laparoscopic surgery). When evaluated on EndoVis2017VOS, our *Student* network reaches an IoU of 83.77%, 12.30% above the state-of-the-art unsupervised AGSD approach, and only 5.84% below the state-of-the-art MF-TAPNet approach.

The proposed unsupervised approach for surgical tool segmentation of optical flow images outperforms state-of-the-art CIS approach by a large margin on EndoVis2017VOS (+16.32% ΔIoU). Most common errors resulted from passive tools like *graspers*, especially when used to hold anatomical structures. Nonetheless, the proposed *learning-from-noisy-labels* strategy showed great robustness to the noise of motion-derived pseudo-

labels, allowing to effectively train the *Student* network: results break-down by tool type (Table 3.7) showed no significant difference in *Student*'s segmentation performance between passive and active surgical tools.
Ablation studies proved that the method is extremely robust to the way *shape-priors* are obtained, with no significant performance difference between using automatically segmented tools from green-screen recordings and *recycled* manual annotations from other datasets. In addition, FUN-SIS showed great robustness to limited *shape-priors* quantity, performing optimally on EndoVis2017VOS even using as few as 51 RoboTool *shape-priors* masks for training. We believe that the loose requirements FUN-SIS has on *shape-priors* type and quantity strongly contribute to justify why it should be regarded as *unsupervised*. Compared to the standard annotations required by weakly-/semi-supervised approaches, *shape-priors* do not need to be generated any time new unlabelled data are collected, as they are agnostic to almost any inter-domain variation and *unpaired* with the endoscopic images. This makes them suitable to be used across different domains as empirically proven for the RoboTool *shape-priors* (used in combination with unlabelled videos of multiple surgical procedures, performed with manual or robotic laparoscopic approaches, involving different tools, recorded with different acquisition systems under significant lighting conditions variations).
Ablation studies highlighted other interesting aspects, such as the benefits of using a log Intersection-over-Union loss when training on noisy pseudo-labels, and the effectiveness of the proposed optical flow augmentation strategy on video object segmentation. Finally, the extensive analysis on pseudo-label noise properties and their impact on neural network training, as well as the proposed *learning-from-noisy-labels* strategy to leverage them, may serve as the base for future work on object segmentation using noisy labels, still largely unexplored.
Despite the satisfying results, the proposed work still presents potential room for improvement:

- the FUN-SIS performance is overall influenced, and bottle-necked, by the quality of the optical flow images, which depends, in turn, on the endoscopic camera resolution and the optical flow estimator. As discussed in Section 5.2, this contributes to the reduced ability of the *Student* network, compared to fully-supervised approaches, to capture fine-scale details. Current research on models for optical flow computation specifically tailored for endoscopic images, as well as the increasing use of high-definition endoscopic cameras, could naturally contribute to improve the effectiveness of the proposed FUN-SIS method. In addition, the quality of the pseudo-labels could be improved using unsupervised post-processing solutions, such as Conditional Random Fields, before being used for *Student* network training. An alternative that could be worth exploring is the use of depth images from stereo-cameras in combination with optical flow images. Similarly to optical flow, depth image domain projection allows to reduce the amount of variability in background and tool appearance compared to the raw endoscopic image domain. This makes it easier to generate pseudo-labels in an unsupervised way. Whenever stereo-camera systems are available (as in several robotic systems nowadays), this combined approach may lead to greater robustness and overall improved segmentation results, than using optical flow only.

- when selecting well-labelled regions through *local* IoU, a great amount of

the available data are currently discarded (49.52% of total available pixels in EndoVis2017VOS experiment), which may impact the generalization ability of the *Student* network. These *uncertainly*-labelled pixels could be exploited in combination with unsupervised strategies (e.g. enforcing consistency of prediction between different augmented views of the same region), and contribute to the *Student* network training;

- the window used to compute the *local* IoU has fixed dimensions and is slid regularly on the masks with fixed width and stride. This approach creates a trade-off between effective training size and quality of the selected regions, requiring to properly tune the window size w, as well as the threshold value ϵ_{IoU}. A more flexible approach, adapting to the varying tool size and location, may be beneficial to improve the quality of the selected regions without excessively reducing the effective training size;

- the *Proxy* network is subjected to strong gradients while training directly on the noisy pseudo-labels, resulting in possible performance oscillations. This can potentially hinder the *Student* network training, if the *Proxy* network training is stopped in a poor weight parameters configuration. This problem could be mitigated by using approaches such as self-ensembling [NMN+20], regularizing *Proxy* network training.

As extensively discussed, FUN-SIS was developed as an unsupervised approach, trainable on a virtually unlimited set of unlabelled data. Nonetheless, an interesting research direction could be exploring its use to guide the selection of particularly challenging samples (like images which provide a high loss value for the Student), in order to focus human annotation effort on few, informative cases.

3.7 Conclusion

In this Chapter we presented FUN-SIS, a Fully-UNsupervised approach for Surgical Instruments Segmentation. The obtained results, almost on par with the ones of fully-supervised solutions, show the value of learning from unlabelled data, exclusively relying on weak prior knowledge, easy to obtain and highly repurposable across various surgical domains. While several research directions can be explored to improve the quality of binary segmentation, as discussed in Section 3.6, the next Chapter builds on top of this work to solve the semantically richer task of instance segmentation.

CHAPTER 4

PAF-IS: a Pixel-wise Annotation Free framework for Instance Segmentation of Surgical Tools

Contents

4.1 Introduction 96
4.1.1 Objective & Contributions 97
4.2 Methodology 97
4.2.1 Tool Instantiation 98
4.2.2 Instance-wise Feature Representation Learning 101
4.2.3 Instance-wise Tool Type Classification 102
4.3 Experimental Set-up 105
4.3.1 Datasets 105
4.3.2 Design Choices & Training Details 106
4.4 Experiments and Results Analysis 107
4.4.1 Tool Instantiation 107
4.4.2 Tool Instance Segmentation 108
4.5 Ablation Studies 109
4.5.1 Tool Instantiation Augmentation Strategy 110
4.5.2 Tool Instantiation Inference Parameters 111
4.5.3 Prototype Labels Number 111
4.6 Discussion 113
4.7 Conclusion 114

4.1 Introduction

As established in the previous Chapter, binary tool segmentation can be learnt from unlabelled data by purely relying on prior knowledge about surgical tool shape and motion. The current Chapter builds on this work to extend the use of prior and complementary knowledge to the instance segmentation problem.
Compared to binary segmentation, instance segmentation significantly increases the value of the extracted information, as it enables to obtain individual tool masks and to simultaneously identify their tool type. Because of its complexity, this task is commonly tackled by fully-supervised Deep Learning approaches [SRKI18a, JCDH19a, KJD+21, KMNA+21]. Such approaches require the availability of pixel-wise semantic and instance labels to train, extremely expensive to collect at a large scale via manual annotation. Indeed, as discussed in the Related Work Chapter, and observable from Table 2.1, research on alternatives to full-supervision has remained confined to the binary segmentation task. We believe that this is due to the rigid problem formalization imposed by common instance segmentation approaches: such approaches do not benefit from the potential availability of binary segmentation masks, as they would still require pixel-wise semantic and instance labels to train. Furthermore, their problem formalization prevents the incorporation of potentially cheaper sources of semantic information, compared to pixel-wise annotations, like binary tool presence labels. Specifically, we define as *frame-wise* those binary tool presence labels describing which tool types are *effectively* visible in each frame; we define as *sequence-wise* those labels indicating which tool types are *potentially* visible in each frame. In robot-assisted surgery, for example, robotic systems can often record which tools are attached to the system [KMNA+21]: however, this information only indicates that a certain tool could be in-use (and visible) at some point of the sequence of frames during which it is attached, but does not guarantee its visibility at a specific frame (therefore a *sequence-wise* visibility). As a generalization, surgical phase and step annotations can provide similar information, when a mapping between phases/steps and tools can be approximately defined (e.g. by knowing which tools are commonly used in each phase/step [PBA+12]). These considerations allow to extend the concept of *sequence-wise* labels to laparoscopic procedures. Examples of *frame-wise* and *sequence-wise* labels are shown in Figure 4.1.

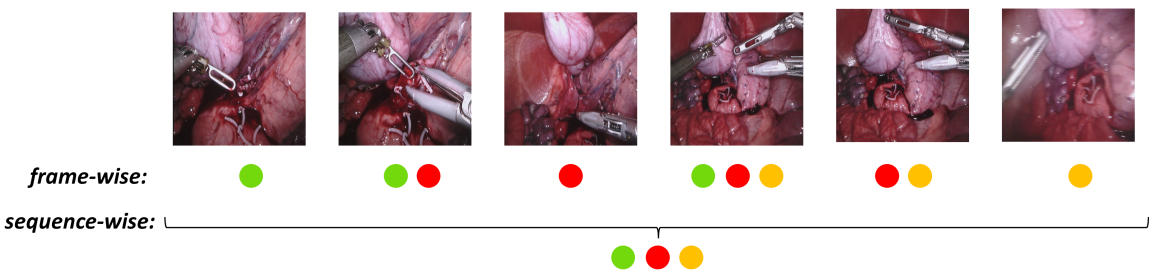

Figure 4.1: Examples of frame-wise and sequence-wise binary tool presence labels for a robot-assisted surgery sequence (each color represents a tool type). All the tools can be attached to the system at the same time, while being visible only in certain frames.

4.1.1 Objective & Contributions

In this Chapter we propose a framework for instance segmentation model training, which embraces the recent progress on unsupervised binary segmentation and the availability of cheap binary tool presence labels, to minimize human annotation effort. Overall, our solution learns to instantiate binary segmentation masks, and to obtain, for each extracted tool, a powerful feature representation using a self-supervised formulation. These instance-wise representations are then used to select a small number of tool instances (*prototype instances*), which are presented to a potential human user for tool type labelling. The gathered information, along with binary tool presence labels (either *frame-wise* or *sequence-wise*), is finally used to train an instance-wise classifier that predicts tool type labels for each tool instance.
At inference time the trained architecture can perform instance segmentation on single frames, by extracting individual tool instances and separately classifying them.
Overall, we make the following contributions:

- we develop an unsupervised approach to learn **tool instantiation from binary segmentation masks** (Figure 4.2, *Tool instantiation*): with no availability of pixel-wise instance labels, we fabricate a pseudo-supervision signal from Connected Component instantiation of the binary masks, and refine it using simple assumptions on instrument positioning in the image space to effectively train an instantiation model;

- we develop an approach for **self-supervised instance-wise feature representation** learning (Figure 4.2, *Instance-wise feature learning*): with no availability of pixel-wise semantic labels, we learn such representations by relying on intrinsic temporal information from video sequences. Specifically, we design a contrastive learning approach based on local instance tracking to draw positive and negative samples. This step allows to obtain powerful instance-wise feature representations, providing the necessary information to solve the final classification step;

- we develop an approach to learn **instance-wise tool type classification with no ground truth labels available** (Figure 4.2, *Instance-wise tool type classification*): the learnt instance-wise feature representations are used to guide the automatic selection of a tiny number of prototype instance tools (as few as 8 in our experiments), displayed to a potential human user for tool type labelling. The gathered information is propagated to the whole training set, allowing to label each training instance with a pseudo-GT tool type label (*prototype label*). This information is combined with the available binary tool presence labels (either *frame-wise* or *sequence-wise*) to solve the instance classification task, exploiting a teacher-student problem formulation.

4.2 Methodology

The proposed PAF-IS framework for Pixel-wise Annotation Free Instance Segmentation explicitly separates the task into three core components: instrument instantiation,

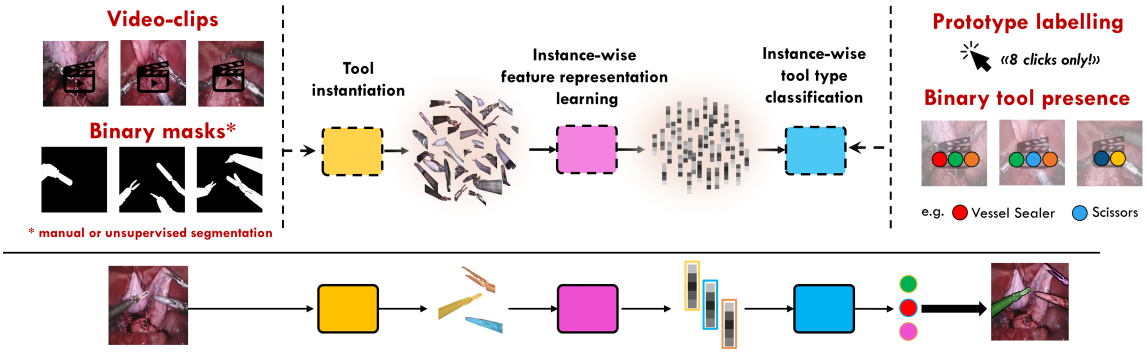

Figure 4.2: Overview of the proposed Pixel-wise Annotation Free framework for Instance Segmentation (PAF-IS). Top: training architecture highlighting the three core steps. Tool instantiation is learnt from binary masks, potentially obtained using recent unsupervised segmentation methods. Instance-wise feature representation learning is performed using a contrastive learning strategy, powered by local temporal tracking. This step allows to extract a feature representation of each tool instance in the training set. Instance-wise tool type classification is performed by incorporating a minimal amount of human-provided information (prototype labels, as few as 8 in our experiments) and cheaply obtainable binary tool presence labels. Bottom: PAF-IS inference architecture.

instance-wise feature learning and instance classification. Differently from standard semantic/instance segmentation approaches, PAF-IS does not require pixel-wise semantic or instance annotation of the training data. Instead, it relies on the availability of binary segmentation masks, which can be cheaply obtained using emerging unsupervised approaches, and binary tool presence labels.
The full framework is presented in Figure 4.2 and detailed below.

4.2.1 Tool Instantiation

Instrument instantiation is here defined as the problem of predicting, from an endoscopic image I, the set of binary masks $\{M_i^{Inst}\}$, with i in $[1, N_{Inst}]$, each one corresponding to an individual instrument visible in the image. When the ground truth instantiation is known, the problem is often formulated as bounding-box prediction [KJD+21, GBSA20]. However, the effectiveness of this approach has been questioned in [KMNA+21], which proposed an alternative solution based on direct regression of instance centroids' position. We here adopt a similar formulation, proving its benefits with respect to bounding-box prediction beyond fully-supervised learning.
The instantiation problem is here formalized as learning the mapping between the image $I \in R^{W \times H \times 3}$ and the displacement field $D \in R^{W \times H \times 2}$, uniquely assigning each tool pixel to an instance. Given a pixel $\mathbf{p} = [p_x, p_y]$, $D|_{\mathbf{p}}$ is equal to the vector $\mathbf{v} = [c_x^i - p_x, c_y^i - p_y]$ if $\mathbf{p}$ belongs to a certain instance i, having its centroid in $[c_x^i, c_y^i]$, or to the null vector $[0,0]$, if $\mathbf{p}$ belongs to the background. Given a training set with known ground truth instantiation D, such mapping can be learnt by an instantiation model, implemented as a neural network, by using a fully-supervised training formulation, as in [KMNA+21]. This can be achieved by optimizing the loss L_I^{FS}, implemented as the pixel-wise distance between the ground truth displacement field D and the instantiation model prediction $\tilde{D}$:

$$L_I^{FS} = |D - \tilde{D}|. \tag{4.1}$$

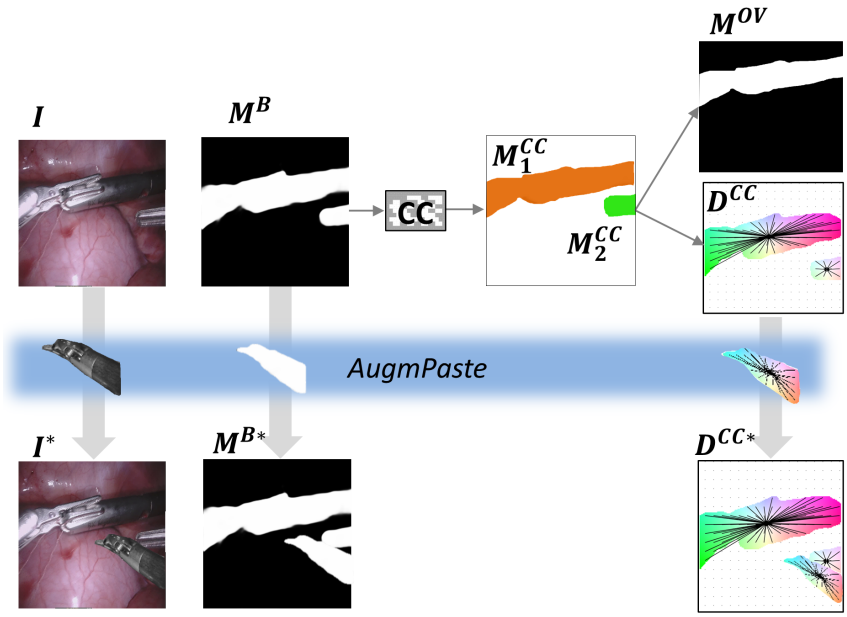

Figure 4.3: Overview of the proposed strategy to generate a pseudo-supervision signal to learn instrument instantiation. Given an image I, its binary mask M^B is instantiated using a Connected Component (CC) algorithm, yielding the set of tool masks $\{M_i^{CC}\}$, with i in $[1, N_{CC}]$. From them, the displacement field D^{CC} and the overlap mask M^{OV} can be automatically obtained. A random tool instance is then selected from the training set, and pasted on I, M^B, D^{CC}, producing their augmented versions I^*, M^{B*}, D^{CC*}.

At inference time, given a new image I and the corresponding predicted displacement field $\tilde{D}$, the set of instance masks $\{M_i^{Inst}\}$ can be easily extracted by identifying the instance centroids, as the pixels where the displacement field converges, and assigning each tool pixel to the centroid pointed by the corresponding displacement vector.

Training: In our case, only the binary mask M^B is known. Without a ground truth instantiation we rely on the assumption that surgeons tend to avoid surgical instruments overlap, in order to reduce the chances of mutual tool occlusions and unwanted tool interactions.
Given an image I and the corresponding binary mask M^B, if tools do not overlap, the instance masks can be obtained by separating the Connected Components (CC) of M^B through standard CV methods like the Spaghetti algorithm [BABG19]. The displacement field D^{CC}, approximating the ground truth D, can then be directly obtained from the set of N_{CC} tool masks $\{M_i^{CC}\}$, with i in $[1, N_{CC}]$, by subtracting each tool pixel position from the centroid $[c_x^i, c_y^i]$ of the corresponding mask M_i^{CC}. While effective in the case of non-overlapping tools, CC labelling systematically fails when tools overlap. In order to mitigate this problem we artificially modify the supervision signal obtained from CC instantiation, as follows:

- **potential overlapping tools identification**: given the set of CC masks $\{M_i^{CC}\}$, M_i^{CC} is considered a potential overlapping instance if it extends across the whole frame (see Figure 4.3 for an example). All pixels corresponding to potential overlapping instances are collected in the binary overlap mask M^{OV}, and discarded from loss computation as described later in this Section;

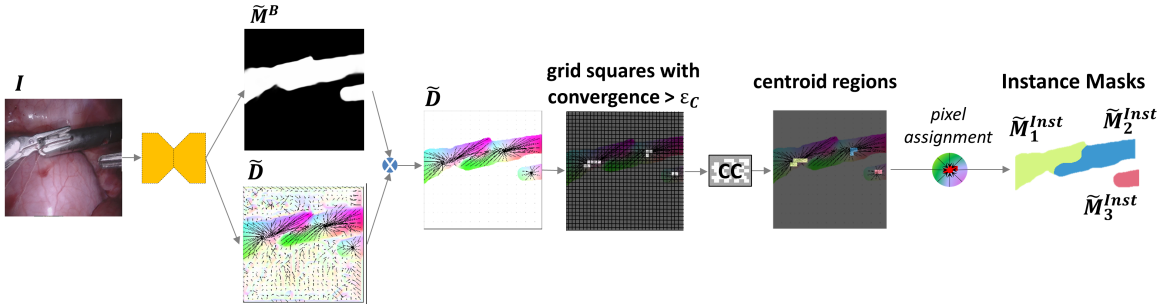

Figure 4.4: Overview of the proposed instantiation strategy. Given an image I the trained instantiation model predicts the masked displacement field $\tilde{D}$. A square grid is then overlapped to $\tilde{D}$, and the squares with high convergence (per-pixel average > ε_C) are then extracted, and separated by Connected Component (CC) labelling, yielding a set of $\tilde{N}_{Inst}$ centroid regions. Each tool pixel is then assigned to the corresponding centroid, yielding the set of instances masks $\{\tilde{M}_i^{Inst}\}$, with i in $[1, \tilde{N}_{Inst}]$.

- **instance pasting augmentation (*AugmPaste*)**: given an image I, its binary mask M^B and its CC displacement field D^{CC}, a random tool instance is selected from a different training sample and pasted on them, yielding the augmented image I^*, the augmented binary mask M^{B*} and the augmented displacement field D^{CC*} (Figure 4.3). This augmentation step allows to artificially simulate the presence of overlapping instances, making up for the discarded instances at the previous step.

Given the image I^*, in addition to the displacement field $\tilde{D}$, we let the instantiation model predict the binary segmentation mask $\tilde{M}^B$, which we multiply by $\tilde{D}$ to ensure that the displacement vector for pixels belonging to the background is a null vector $[0,0]$. For simplicity, we keep the notation $\tilde{D}$ to refer to the result of such product.
Given the image I^*, the corresponding network predictions $\tilde{D}$ and $\tilde{M}^B$, the binary mask M^{B*}, the displacement field D^{CC*} and the overlap mask M^{OV}, the instantiation model is trained by optimizing the loss L_I:

$$L_I = |D^{CC*} - \tilde{D}|(1 - M^{OV}) + L_{CE}(M^{B*}, \tilde{M}^B), \tag{4.2}$$

where L_{CE} is a standard pixel-wise cross-entropy loss.

Inference: given an image I and the trained instantiation model, the predicted displacement field $\tilde{D}$ must me mapped to the set of instance masks $\{\tilde{M}_i^{Inst}\}$, with i in $[1, \tilde{N}_{Inst}]$, and $\tilde{N}_{Inst}$ being the number of predicted instances in a frame. While for the ground truth displacement field D each tool pixel vector points exactly to the corresponding centroid pixel, this is not guaranteed for the predicted $\tilde{D}$. Therefore we define as *centroids* the regions of $\tilde{D}$ with a high rate of displacement vectors convergence. Practically, we overlap a square grid to $\tilde{D}$ and compute, for each square, the per-pixel average number of vectors pointing inside it. If such number is above a predefined threshold ε_C, the square is considered a centroid square. Connected squares are grouped together, to yield the set of centroid regions $\{c_i\}$, with i in $[1, \tilde{N}_{Inst}]$. The instance masks can then be extracted by assigning each tool pixel $\mathbf{p}$ to the centroid c_i closest to the point identified by $\mathbf{p} + \tilde{D}|_{\mathbf{p}}$. This yields the set of predicted instance masks $\{\tilde{M}_i^{Inst}\}$, with i in $[1, \tilde{N}_{Inst}]$ (Figure 4.4). In our framework, the predicted instance masks are subsequently used to learn instance-wise feature representations, as now discussed.

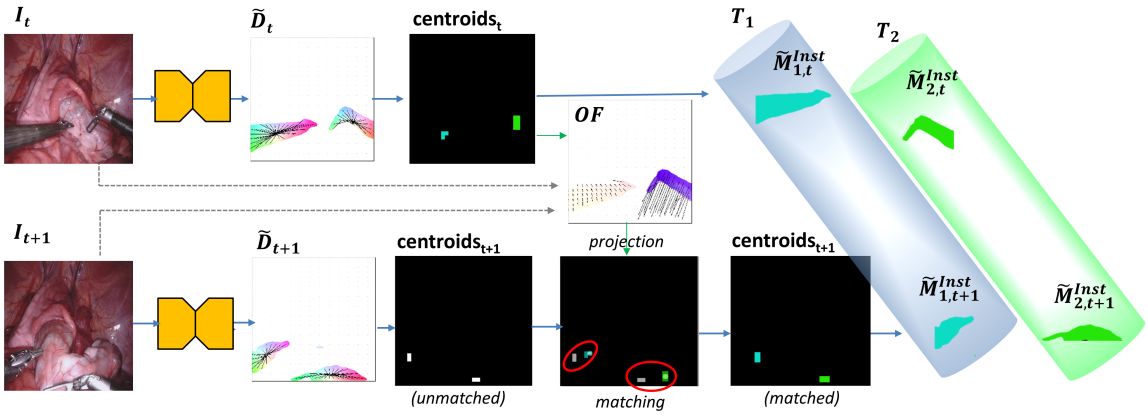

Figure 4.5: Overview of the tracking strategy used to generate positive samples for contrastive learning. Given two consecutive frames, centroids at time t, obtained from the displacement field D_t are mapped to the I_{t+1} space using optical flow OF, computed between the two images I_t and I_{t+1}. The projected centroids are then matched to the ones obtained from the displacement field D_{t+1}. This allows to build the set of tubes $\{T_i\}$, with i in $[1, \tilde{N}_{Inst}]$. Tubes are progressively grown by repeating this process for consecutive frames.

4.2.2 Instance-wise Feature Representation Learning

In the absence of pixel-wise semantic labels, we rely on self-supervision to learn robust and meaningful feature representations of each tool instance, tailored for the instance segmentation task. The problem of self-supervised representation learning has been often addressed by means of contrastive learning in literature [JBZ+20]. While general contrastive learning approaches usually learn global frame-level feature representations, we find this formulation to be ill-posed for the instrument segmentation problem, as it lacks the spatial granularity necessary to discriminate between different instances. Therefore we design an instance-level contrastive learning approach, exploiting the unsupervised instantiation described above and intrinsic temporal information from video sequences. Given an image I and the set of instance masks $\{\tilde{M}^{Inst}_i\}$ predicted by the instantiation model, we want to map each instance to a feature vector F_i, capturing its semantic content. We obtain feature vectors using a feature extractor model implemented using a standard ResNet-50 architecture. Specifically, for each instance, we pass I through the model and multiply the intermediate feature maps by $\tilde{M}^{Inst}_i$, resized to match their dimensions, to obtain the corresponding instance-wise feature vector F_i. Then, given a feature representation F_i, intrinsic temporal information from the video sequence is used to draw positive and negative examples for contrastive loss computation. Specifically:

- positive examples $\{F^+_i\}$ are sampled from the instance tube T_i, built from the frame-by-frame tracking of the instance i. Such tracking is described in Figure 4.5. Given the consecutive images I_t and I_{t+1}, and their corresponding sets of instrument instances, tracking is solved by projecting the centroids of I_t into I_{t+1} space using the optical flow OF, computed between I_t and I_{t+1}. Each I_t centroid is then matched to the closest I_{t+1} centroid. Optical flow projection allows to robustly handle tool movements between consecutive frames, reducing the chances of wrong matching;

- negative examples $\{F^-_i\}$ can be sampled either from different tubes belonging to the same frame, or from tubes far apart in time.

The feature extractor network is then trained by optimizing the loss L_F between $\{F_i^+\}$ and $\{F_i^-\}$:

$$L_F = L_{SCL}(\{F_i^+\}, \{F_i^-\}), \tag{4.3}$$

where L_{SCL} is the Supervised Contrastive Loss formulation proposed in [KTW+20], with each instance tube treated as a separate class. The learnt feature representations are exploited in the next step for classifier training.

4.2.3 Instance-wise Tool Type Classification

Given the set of available tool type classes $\{S_i\}$, with i in $[1, N_{cls}]$, a classifier model must now be trained to learn the mapping between instance-wise features and class labels from that set. In the absence of pixel-wise semantic labels, we rely on binary tool presence labels to solve this task. However, in order to leverage this information, each tool instance must be matched to a binary tool presence label. To this aim, we inject a minimal amount of human knowledge, specifically collected to maximize its information content and exploit it to solve the matching task.
Specifically, we automatically select a tiny number of highly representative instances (*protoype* instances) and ask a human user to label them. The gathered information is then used to match binary tool presence labels and instances, providing an effective supervision signal for classifier training. The two steps are now detailed.

Prototype labelling: given the complete set of learnt features for all the instances in the training set, unsupervised clustering is applied. In our experiments we make use of the standard K-Means++ clustering algorithm [AV06], with the number of clusters N_{km} regarded as an hyper-parameter. The N_{km} instances corresponding to the clusters' centroids are defined as *prototype instances*. A human user would now be required to assign a label S^P from the set $\{S_i\}$ to each prototype instance. In order to propagate the prototype instance labels to the rest of the training instances, we require all instances belonging to the same cluster to share the same semantic label S^P. Figure 4.6 provides a visualization the of prototype instance labelling process, and of the result of prototype labels propagation. In principle, a number of clusters N_{km} equal to N_{cls}, the total number of tool type classes available, is sufficient to correctly label the whole training dataset, and potentially to directly deploy the instance segmentation model: given an unseen image I and a predicted tool instance mask $\tilde{M}_i^{Inst}$ from that image, inference would then be performed by extracting the corresponding feature vector F_i and associating it to the prototype label S^P of the cluster closest to F_i in the feature space. However, in practice, as the feature learning step is imperfect, the prototype instance label propagation is also imperfect. Nonetheless, we show that the information provided by prototype labels can be used to match binary tool presence labels and instances, providing an effective supervision signal for classifier training.

Binary tool presence labels incorporation: let us consider the set of binary tool presence labels $\{S_i^W\}$ with i in $[1, N_W]$, subset of the set of tool type labels $\{S_i\}$, associated to a certain frame. As discussed in the Introduction Section, this information can be defined as *frame-wise*, if the labels indicate which tool types are *effectively* visible in the frame, or *sequence-wise*, if they indicate which tool types are visible at some point in the sequence the frame belongs to, but not necessarily in such frame. Binary tool presence information, either *frame-wise* or *sequence-wise*, does not provide tool localisation information, and is therefore defined as *weak* with respect to the segmentation task. While cheaply

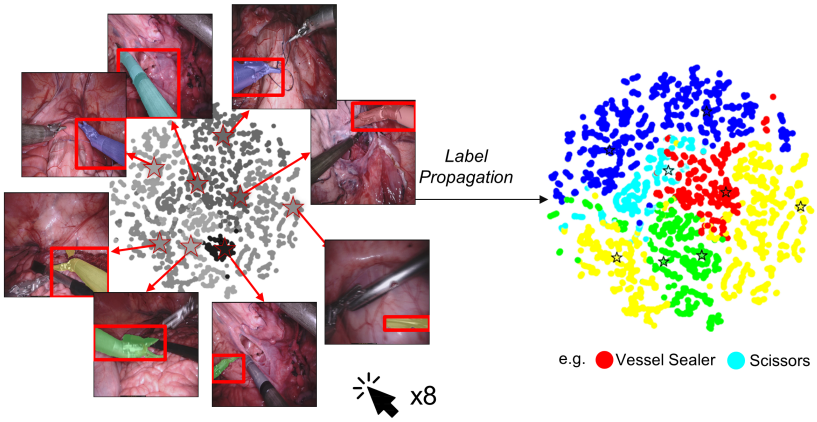

Figure 4.6: Left: visualization of learnt feature representations of the EndoVis 2017 [ASK+19] training set instances, clustered (N_{km} equal to 8) and projected in the 2D space using t-SNE algorithm [VdMH08]. Each instance point is colored in a different shade of grey to represent the cluster id. Prototype instance features are marked with ⋆, and the corresponding masks are overlaid on the frame and highlighted by a bounding-box, to facilitate their labelling by a human user. The color of the mask overlays represents the ground truth tool type that the user would assign. Right: prototype instance labels propagation to the training set. Each instance-wise feature projection is colored accordingly to its prototype label, assigned via propagation from the prototype instance.

obtainable, such weak labels are often overlooked by segmentation approaches, as they pose several challenges:

- differently from pixel-wise semantic labels, weak labels are not directly matched to a specific instance, making them hard to digest for standard segmentation approaches, whose training is based on pixel-wise annotations;
- depending on the system/annotation protocol used to collect the information, the presence of multiple instances of the same tool type may not be recorded. In the Cholec80 dataset [TSM+16], for example, *frame-wise* binary tool presence labels do not keep track of multiple tool instances;
- for *sequence-wise* labels it is quite common that tool type labels do not reflect which tools are effectively visible in the image. In the case of robotic surgery, for example, tools are attached beforehand to the robotic system, potentially remaining unused for relatively long periods of time. Similarly for phases, certain tools, like the ones used for coagulation, may be linked to every phase of a procedure, while being visible only for small amounts of time.

In order to make effective use of such information, each tool instance in a frame must be matched to a weak label from the set $\{S_i^W\}$ associated to that frame. Once the matching is found, a classifier model can be trained on the matched labels. In practice, the binary tool presence labels softly constrain the training of the classifier, providing a reduced set of tool type labels among which the ground truth one for each instance is to be found.

Let us consider an image I, the sets of instance masks, features and prototype labels $\{\tilde{M}_i^{Inst}\}, \{F_i\}, \{S_i^P\}$, with i in $[1, \tilde{N}_{Inst}]$, and the set of weak labels $\{S_i^W\}$, with i in $[1, N_W]$ associated to I. Mining such weak labels requires finding the function ξ, matching the set of $\tilde{N}_{Inst}$ features to the set of N_W weak labels. However, in the most general case, such transformation is:

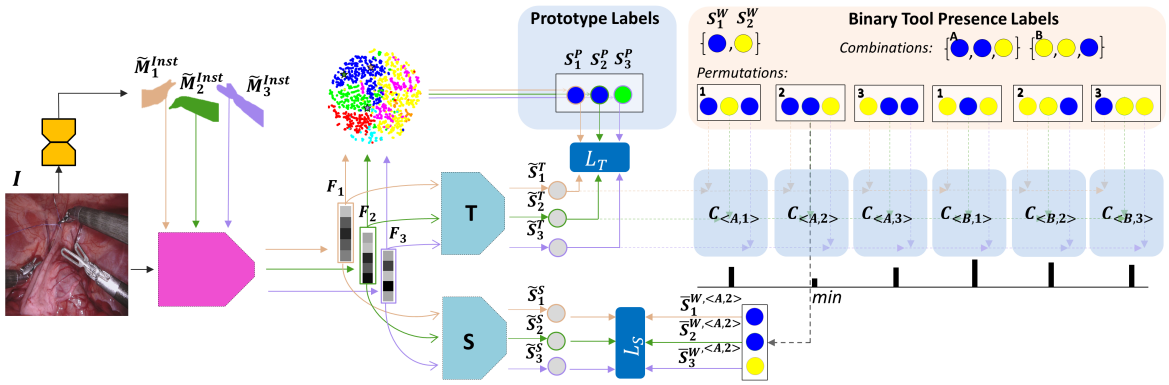

*Figure 4.7: Overview of the proposed weakly-supervised instance classification module. Given an image I, the corresponding set of instance-wise features $\{F_i\}$, with i in $[1, \tilde{N}_{Inst}]$, is obtained from the instance masks $\tilde{M}_i^{Inst}$. Each feature is mapped to the corresponding prototype label S_i^P, which, as shown in this case, does not necessarily correspond to the ground truth label. Each feature is also independently passed through the Teacher (**T**) and Student networks (**S**), yielding the predicted probabilities $\tilde{\mathbf{P}}_\mathbf{i}^\mathbf{T}$, $\tilde{\mathbf{P}}_\mathbf{i}^\mathbf{S}$, and the corresponding predicted labels $\tilde{S}_i^T$, $\tilde{S}_i^S$ (for the sake of readability only the latter are shown in the picture). **T** is trained optimizing the loss L_T computed using the prototype labels $\{S_i^P\}$. Simultaneously, **T** predictions are used to compute the assignment costs $C_{\langle i_C, i_P \rangle}$ for each i_P permutation of each i_C combination of the weak labels $\{S_i^W\}$, with i in $[1, N_W]$. The ordered set $[\bar{S}_i^W]$, with i in $[1, \tilde{N}_{Inst}]$, corresponding to the minimum assignment cost, is used to compute the loss L_S for Student network optimization.*

- *non injective*, as there could be multiple instances sharing the same tool label S_i^W;
- *non surjective*, as a certain tool label S_i^W may not be present in a specific frame.

This implies that given the set of $\tilde{N}_{Inst}$ tool instances in a frame, different combinations of $\tilde{N}_{Inst}$ elements of the N_W weak labels are plausible. To simplify the problem, and avoid degenerate solutions, we assume that if the number of instances $\tilde{N}_{Inst}$ in a frame is equal or smaller than the number of labels, every instance is assigned to a different label. Specifically, we identify the set of plausible weak labels combinations as follows:

- if $\tilde{N}_{Inst} < N_W$, all the possible combinations of $\tilde{N}_{Inst}$ elements of the N_W labels are plausible;
- if $\tilde{N}_{Inst} == N_W$, we assume that the set of N_W labels is the only plausible combination;
- if $\tilde{N}_{Inst} > N_W$, all the possible combinations with repetitions of $\tilde{N}_{Inst}$ elements of the N_W labels are plausible.

Among the set of plausible weak label combinations, the correct label combination must be identified, and the matching between each instance and each weak label in such combination must be determined. This could be achieved by associating to each i_P permutation of each i_C plausible combination of the weak labels an assignment cost $C_{\langle i_C, i_P \rangle}$. Each couple $\langle i_C, i_P \rangle$ yields an ordered set of weak labels $[S_i^{W,\langle i_P, i_C \rangle}]$, with i in $[1, \tilde{N}_{Inst}]$. Among them, the ordered set minimizing the assignment cost could be selected and used for the classifier training.
To solve this problem we propose a teacher-student approach (Figure 4.7), exploiting the

knowledge gathered from the prototype labels. Teacher and Student are two identical classifiers that map a feature vector F_i to the vectors $\tilde{\mathbf{P}}_\mathbf{i}^\mathbf{T}$, $\tilde{\mathbf{P}}_\mathbf{i}^\mathbf{S}$, respectively. $\tilde{\mathbf{P}}_\mathbf{i}^\mathbf{T}$, $\tilde{\mathbf{P}}_\mathbf{i}^\mathbf{S}$ represent the predicted probability of the instance to belong to each of the N_{cls} classes, according to Teacher and Student, respectively. From $\tilde{\mathbf{P}}_\mathbf{i}^\mathbf{T}$, $\tilde{\mathbf{P}}_\mathbf{i}^\mathbf{S}$ the class with the highest probability $\tilde{S}_i^T$, $\tilde{S}_i^S$ is regarded as the predicted label. The Teacher network is trained to map each feature F_i to the corresponding prototype label S_i^P, by optimizing the instance-wise classification loss L_{T_i}:

$$L_{T_i} = L_{CE}(\tilde{\mathbf{P}}_\mathbf{i}^\mathbf{T}, S_i^P). \tag{4.4}$$

For each couple $\langle i_C, i_P \rangle$, its assignment cost $C_{\langle i_C, i_P \rangle}$ can then be computed as the average cross-entropy loss between the predicted probabilities $[\tilde{\mathbf{P}}_\mathbf{i}^\mathbf{T}]$ and the weak labels $[S_i^{W,\langle i_P, i_C \rangle}]$, corresponding to that couple, as follows:

$$C_{\langle i_C, i_P \rangle} = \frac{1}{\tilde{N}_{Inst}} \sum_{i=1}^{\tilde{N}_{Inst}} L_{CE}(\tilde{\mathbf{P}}_\mathbf{i}^\mathbf{T}, S_i^{W,\langle i_C, i_P \rangle}). \tag{4.5}$$

The ordered set of weak labels $[\bar{S}_i^{W,i_P,i_C}]$, corresponding to the couple $\langle i_C, i_P \rangle$ minimizing the assignment cost, is selected. The Student network is then trained by optimizing the instance-wise classification loss L_{S_i}, between the predicted probabilities $[\tilde{\mathbf{P}}_\mathbf{i}^\mathbf{S}]$ and the matched weak labels $[\bar{S}_i^{W,i_P,i_C}]$:

$$L_{S_i} = L_{CE}(\tilde{\mathbf{P}}_\mathbf{i}^\mathbf{S}, \bar{S}_i^{W,\langle i_C, i_P \rangle}). \tag{4.6}$$

In practice, the Teacher network applies the knowledge gathered from the prototype labels to identify the correct ordered set of weak labels used for Student training. Doing so, the Teacher approximates the function ξ, matching each of the $\tilde{N}_{Inst}$ tool instances to a weak label from the set $\{S_i^W\}$.
This general framework applies to both *frame-wise* and *sequence-wise* binary tool presence labels. In the case of *frame-wise* labels, ξ becomes surjective, significantly reducing the space of possible solutions and facilitating the matching.

4.3 Experimental Set-up

The proposed framework was validated on the MICCAI 2017 and 2018 EndoVis Robotic Instrument Segmentation Challenge datasets. The two datasets are now introduced (Section 4.3.1), together with the specific design choices and training details (Section 4.3.2).

4.3.1 Datasets

EndoVis2017 [ASK+19]: the original challenge dataset consists of 10 video clips, resampled at a frame rate of 1 frame-per-second, of abdominal porcine procedures, performed using da Vinci robotic system. Each clip contains 300 high-resolution frames (1024 × 1280). During the challenge 8x 225 frames were released for training, while the remaining 8x 75 frames and two additional clips were held out by the organizers for testing. A total of 7 tool classes are present in the dataset. We provide results on this dataset according to the same evaluation protocol as [SRKI18a], by performing 4-fold cross-validation on the 8x 225 released training data (regrouped in 4 splits). We

report the average metric over the 4 splits, for direct comparison with state-of-the-art approaches.

EndoVis2018 [AKB+20]: the original challenge dataset contains 19 video clips, resampled at a frame rate of 1 frame-per-second, of abdominal porcine procedures, performed using da Vinci robotic system. Each video contains a total of 300 high-resolution frames (1024 × 1280). During the challenge 15 clips were released for training, while the remaining clips were held out by the organizers for testing. The dataset was originally annotated for anatomy and tool-part segmentation, and did not feature instrument type labels. [GBSA20] annotated with pixel-wise tool type labels 149 frames for each of the 15 training clips, and split them into a training set consisting of 11 clips, and a validation set containing the remaining 4 clips. The same 7 tool classes from EndoVis2017 dataset were used. We provide results on this dataset according to the same evaluation protocol as [GBSA20], by training on the 11 training clips, and validating on the remaining 4 clips.

As the proposed PAF-IS approach requires binary instrument masks to train, we provide results using both manually annotated binary masks and automatically segmented masks generated using the unsupervised FUN-SIS approach [SRDM+23]. The mean binary IoU for the FUN-SIS approach on the EndoVis2017 and EndoVis2018 datasets is equal to 83.7% and 81.3%, respectively.

Frame-wise binary tool presence labels were automatically generated for each frame as the unique pixel-wise semantic labels present in the corresponding ground truth masks. *Sequence-wise* binary tool presence labels were also automatically generated, by considering each video clip in the datasets as a sequence, and assigning to each clip, as *sequence-wise* labels, the full set of unique semantic labels present in the ground truth masks of all the frames in the clip. For 46.12% of the frames in the EndoVis2018 dataset the *sequence-wise* labels do not correspond to the *frame-wise* labels (40.72% for EndoVis2017 dataset), i.e., for a certain frame, its *sequence-wise* labels contain at least a tool type which is not visible in it (but which is present at some point in the clip it belongs to).

4.3.2 Design Choices & Training Details

Tool instantiation: the instantiation model is implemented as a U-Net architecture with SegFormer encoder [XWY+21], available from the *Segmentation Models* library in PyTorch. Training was carried out for 60 epochs using the Adam optimizer with a learning rate equal to 1e-3 and a batch size of 32, applying standard photometric and geometric augmentations from the *Albumentation* library to the original images, resized to a 256 × 256 resolution. During inference, centroids were selected by overlapping the predicted displacement field with a square grid of 32 × 32 resolution (i.e. each grid square of 8 × 8 pixel dimension); a threshold ε_C of 5 was used to select centroid squares (i.e. squares with a per-pixel average of at least 5 displacement vectors pointing at them were selected as centroids). The impact of grid resolution and threshold value is investigated in Section 4.5.

Instance-wise feature representation learning: the feature extractor network is implemented as a ResNet-50 architecture. Each instance mask is multiplied by the output of the $conv3\_4$ layer. Instance-wise features are obtained by applying a global average pooling to the output of the $conv5\_3$ layer, having 2048 feature channels.

Training was carried out for 80 epochs using the Adam optimizer with a learning rate equal to 5e-5 and a batch size of 64, applying standard photometric and geometric augmentations to the original images, resized to a 512×512 resolution. For the contrastive loss L_{SCL} a temperature factor equal to 0.1 was used.
Instance-wise tool type classification: for the main experiments (Section 4.4.2), K-Means++ clustering algorithm was applied with a total number of cluster N_{km} equal to 8 (therefore 8 instances were required to be labelled by a potential human user). While in the real scenario such assignment would be performed by a human operator, as discussed in Section 4.5, it was here automatically performed by associating to each prototype instance the semantic label of the ground truth instance of the same frame having the maximum overlap according to the Intersection-over-Union metric.
The classification networks (Teacher, Student) were implemented as a 2-layer fully-connected network, with intermediate feature size of 512 and batch normalization. Training was carried out for 40 epochs using the Adam optimizer with a learning rate equal to 1e-4 and a batch size of 128, applying standard photometric and geometric augmentations to the original images, resized to a 512×512 resolution.

4.4 Experiments and Results Analysis

We now present the experimental validation of the proposed PAF-IS framework, and compare it with state-of-the-art approaches. Tool instantiation results and complete instance segmentation results are separately presented in Sections 4.4.1 & 4.4.2, respectively.

4.4.1 Tool Instantiation

In order to analyze tool instantiation quality, we evaluate results according to a class-agnostic Average-Precision metric, computed for two values of threshold Intersection-Over-Union (IoU): AP@0.5 (50%), AP@0.7 (70%). We present results obtained by our unsupervised approach using, as binary masks, both manual annotations (PAF-IS CC_M) and unsupervised FUN-SIS predictions (PAF-IS CC_F). In addition, we report results for the instantiation model trained in a fully-supervised manner on the ground truth displacement field (PAF-IS GT). As, to the best of our knowledge, no other work has previously attempted unsupervised instantiation of binary tool masks, we compare our solution against a Mask-RCNN baseline, trained under the same fully-supervised (MRCNN GT) and unsupervised modalities (MRCNN CC_M, MRCNN CC_F). However, as Mask-RCNN is an anchor-based approach, the local masking for automatically identified overlapping tools (M^{OV}, described in Section 4.2.1), is not easily implementable, and would require substantial architectural modifications which are beyond the scope of this work. Therefore we limit the augmentation strategy for unsupervised Mask-RCNN experiments to instance pasting, described in Section 4.2.1.

Results presented in Table 4.1 show how our proposed solution outperforms Mask-RCNN across both datasets and for all the three training modalities. A similar result for the fully-supervised training modality was already presented in [KMNA+21]. These experiments highlight the benefits of tool instantiation based on direct centroid regression, beyond full-supervision, for the unsupervised setting. Indeed, the unsupervised PAF-IS

solution using binary annotated masks (PAF-IS CC_M) closely follows the fully-supervised one (PAF-IS GT), with an average gap of -Δ3.3% AP@0.5 across the two datasets. In addition, the greatest performance gap between PAF-IS and Mask-RCNN is found when using FUN-SIS masks to train (CC_F): +Δ17.5% AP@0.5 and +Δ11.15% AP@0.7, in the EndoVis 2017 dataset. This result shows how our solution is particularly suitable to handle a noisy supervision signal. Finally, the performance gap between PAF-IS CC_F and PAF-IS CC_M is significantly smaller for the AP@0.5 metric (-Δ4.48% on average across the two datasets) compared to the AP@0.7 metric (-Δ9.74%). This can be attributed to the lower quality of FUN-SIS binary segmentation masks, causing a performance drop when a high IoU threshold is used: the lower 50% IoU threshold, instead, being less affected by possible inaccuracies in the binary segmentation masks, highlights the high instantiation quality.

Superv.	Method	EndoVis			
		2017		2018	
		AP@0.5	AP@0.7	AP@0.5	AP@0.7
GT	MRCNN	76.11	61.87	75.01	63.12
	PAF-IS	**88.40**	**72.12**	**78.57**	**66.00**
CC_M	MRCNN	71.26	55.98	73.99	60.04
	PAF-IS	**85.36**	**63.70**	**75.92**	**61.08**
CC_F	MRCNN	63.81	44.99	62.48	42.31
	PAF-IS	**81.31**	**56.14**	**71.01**	**49.17**

Table 4.1: Tool instantiation results for the proposed PAF-IS approach and Mask-RCNN on EndoVis 2017 and 2018 datasets, trained according to three modalities: fully-supervised (GT) and unsupervised using Connected Component labelling of manually annotated masks (CC_M) and FUN-SIS predicted masks (CC_F).

4.4.2 Tool Instance Segmentation

In order to evaluate instance segmentation results, and compare them with other state-of-the-art segmentation approaches, we adopt the commonly used IoU EndoVis challenge metric defined in [GBSA20]. It is worth noticing that such metric treats the segmentation problem as pixel-wise classification, without providing information about instantatiation quality. Table 4.2 reports the results of our PAF-IS framework and several state-of-the-art solutions. For each method the table highlights the type of supervision used for training. State-of-the-art approaches are all trained in a fully-supervised way using pixel-wise semantic annotations (S), in combination with pixel-wise instance annotations for instance segmentation methods (I). Our PAF-IS framework does not require pixel-wise semantic or instance annotations to train, relying instead only on prototype instance labels (P) - 8 for the experiments reported in this Table - and weak labels, in the form of *frame-wise* (FW) or *sequence-wise* (SW) tool presence labels (results for both modalities are reported). In addition, PAF-IS can be trained using manually annotated binary masks (B) if available, or rely on the predictions of the unsupervised FUN-SIS approach (results for both modalities are also reported).

Results presented in Table 4.2 show that our PAF-IS approach outperforms fully-supervised and semi-supervised solutions adopting a semantic segmentation problem formulation (Ternaus, MF-TN, DMF-TN), despite not requiring any pixel-wise

Method	Supervision Type						EndoVis	
	Pixel-wise			Weak			2017	2018
	S	I	B	P	FW	SW		
Ternaus[SRKI18a]	✓						35.27	/
MF-TN† [JCDH19a]	✓						37.35	/
DMF-TN† [ZJG+20]	$\checkmark_{30\%}$						**45.83**	/
DMF-TN† [ZJG+20]	$\checkmark_{20\%}$						43.71	/
DMF-TN† [ZJG+20]	$\checkmark_{10\%}$						33.64	/
M&C†‡ [KMNA+21]	✓	✓					**65.70**	/
ISI-Net† [GBSA20]	✓	✓					55.62	73.03
MRCNN[KJD+21]	✓	✓					42.28	/
Tra-SeTr† [ZJH22]	✓	✓					60.04	**76.20**
PAF-IS			✓	$\checkmark_{0.3\%}$			43.86	56.62
PAF-IS			✓	$\checkmark_{0.3\%}$	✓		**53.73**	63.38
PAF-IS			✓	$\checkmark_{0.3\%}$		✓	52.64	**63.57**
PAF-IS				$\checkmark_{0.3\%}$			30.47	54.08
PAF-IS				$\checkmark_{0.3\%}$	✓		**45.86**	**58.03**
PAF-IS				$\checkmark_{0.3\%}$		✓	42.41	57.75

Table 4.2: Instance segmentation results for the proposed PAF-IS approach, state-of-the-art methods on EndoVis 2017 and 2018 datasets. Supervision signals used by each approach are reported: pixel-wise semantic labels (S), pixel-wise instance labels (I), required by fully-supervised instance segmentation approaches, pixel-wise binary segmentation masks (B, for PAF-IS, if not checked FUN-SIS predicted masks are used), prototype labels (P, 8 labels in total in these experiments, ∼0.3% of total training instances), frame-wise tool presence labels (FW) and sequence-wise tool presence labels (SW). † methods using temporal information at inference time. ‡ methods using additional tool-part annotations for training.

annotation. On the EndoVis 2017 dataset our solution also outperforms a standard Mask-RCNN (MRCNN), trained on manually annotated segmentation masks and bounding-boxes for ground truth instantiation. In addition to pixel-wise semantic and instance annotations, the solutions outperforming our PAF-IS approach also rely on temporal information during inference (†) and additional tool-part segmentation annotations (‡). It is worth noticing that temporal modelling is a natural extension for PAF-IS, as tool tracking information is already extracted as part of the instance-wise feature learning step. Qualitative results are shown in Figures 4.12 & 4.13 at the end of the Chapter.

4.5 Ablation Studies

In order to provide a deeper insight into the PAF-IS framework, we now present and discuss ablation studies on three critical design choices: the augmentation strategy for tool instantiation, the inference parameters for tool instantiation and the number of prototype labels required for instance classification training.

4.5.1 Tool Instantiation Augmentation Strategy

In order to train the displacement network for instrument instantiation, a pseudo-supervision signal is generated from the binary masks using a Connected Component algorithm. Such signal is subsequently refined by 1) preventing training on potentially overlapping instances (OV) and 2) pasting random tool instances (PS) to artificially simulate the case of overlapping instances (Section 4.2.1).
Tables 4.3 & 4.4 provide results of an ablation study exploring different combinations of the two augmentation strategies. Such results prove the effectiveness of the two augmentation strategies, and of their simultaneous use. In the case of binary annotated masks, instance masking (OV) provides an average improvement of +Δ4.46% AP@0.5 and +Δ1.08% AP@0.7 across the two datasets, compared to the setting where no augmentation is used; instance pasting (PS) provide an average improvement of +Δ3.03% AP@0.5 and +Δ1.56% AP@0.7; the two strategies combined provide an average improvement of +Δ7.02% AP@0.5 and +Δ5.55% AP@0.7. On the EndoVis 2018 dataset, paste augmentation appears less effective: this could be due to the fact that several frames in it present at least 4 separate tool instances, making the additional pasting redundant, and potentially detrimental as frames can become too cluttered.

Augm.		EndoVis			
OV	PS	2017		2018	
		AP@0.5	AP@0.7	AP@0.5	AP@0.7
		74.85	56.585	71.56	58.08
✓		77.74	54.82	**77.58**	62.00
	✓	81.91	59.82	70.54	57.98
✓	✓	**85.35**	**63.70**	75.92	**62.08**

Table 4.3: Results of the ablation study on unsupervised instrument instantiation from manually annotated binary masks, highlighting the separate and combined impact of: masking of potentially overlapping instances (OV) and pasting of random tool instances (PS).

Augm.		EndoVis			
OV	PS	2017		2018	
		AP@0.5	AP@0.7	AP@0.5	AP@0.7
		67.82	47.86	65.23	43.94
✓		71.80	45.69	**71.91**	48.99
	✓	72.41	49.42	67.99	47.12
✓	✓	**81.31**	**56.14**	71.01	**49.16**

Table 4.4: Results of the ablation study on unsupervised instrument instantiation from FUN-SIS predicted binary masks, highlighting the separate and combined impact of: masking of potentially overlapping instances (OV) and pasting of random tool instances (PS).

4.5.2 Tool Instantiation Inference Parameters

In order to obtain instance masks, a square grid is overlapped to the predicted displacement field; centroid squares are then selected as the ones whose per-pixel average of vectors pointing inside them is greater than the threshold value ε_C. The grid resolution (equal to 32×32 in our main experiments) and the threshold ε_C (equal to 5 in our main experiments) regulate the trade-off between precision and recall of the obtained instance masks. We experimentally evaluate the impact of the two parameters by varying them in a grid-like manner, with grid resolution in $[8, 16, 32, 64, 128]$ and ε_C in $[1, 3, 5, 7, 10]$. Their different combinations are used to obtain instance masks from the same displacement fields. The AP@0.5 between the obtained masks and the ground truth instances is reported in Figure 4.8 for both the EndoVis2017 and EndoVis2018 datasets.

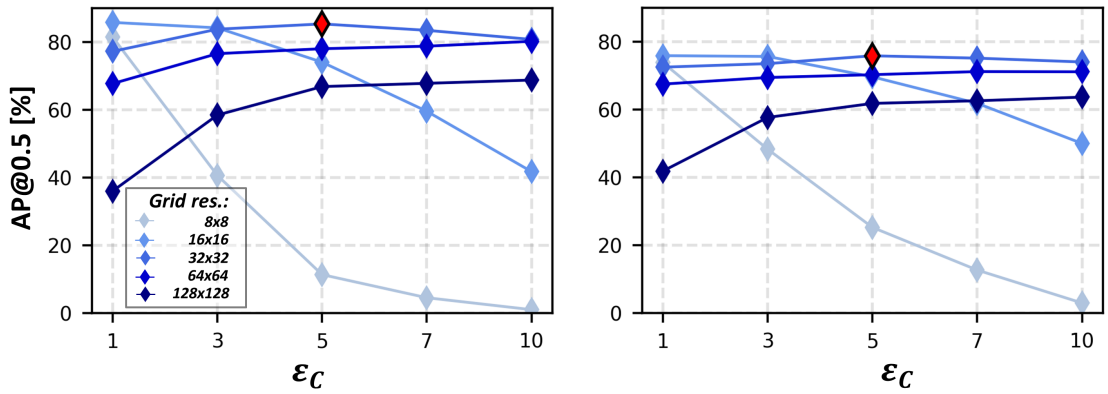

Figure 4.8: Impact of grid square resolution and threshold value ε_C on the tool instantiation quality for the EndoVis2017 dataset (left) and EndoVis2018 dataset (right). The combination used in our main experiments is highlighted in red.

The presented results, together with the qualitative results shown in Figure 4.9, clearly highlight the impact of the two parameters. For intermediate grid resolution values (32×32, 64×64), the impact of ε_C is minimal. However, as the grid solution decreases (16×16, 8×8), an high value of ε_C negatively affects the quality as instantiation, as the average convergence rate on large squares tends to be lower. This can be also observed from the qualitative instantiation results shown in Figure 4.9, top-right, where no candidate squares reach the threshold. Vice-versa, high grid resolution values (128×128) tend to be more negatively affected by a low ε_C, as it leads to the identification of many false positive centroids (instantiation results from Figure 4.9, bottom-left).

4.5.3 Prototype Labels Number

In PAF-IS, the Teacher network is required to gather knowledge from the prototype labels, in order to be able to identify the correct ordered sets of weak labels used for Student training. Prototype labels, therefore can have a crucial influence on the quality of instance classification. In addition, they represent the only piece of human-sourced information necessarily required by PAF-IS for training. Therefore we now present, in Table 4.5, the impact on the segmentation performance, of the number of clusters

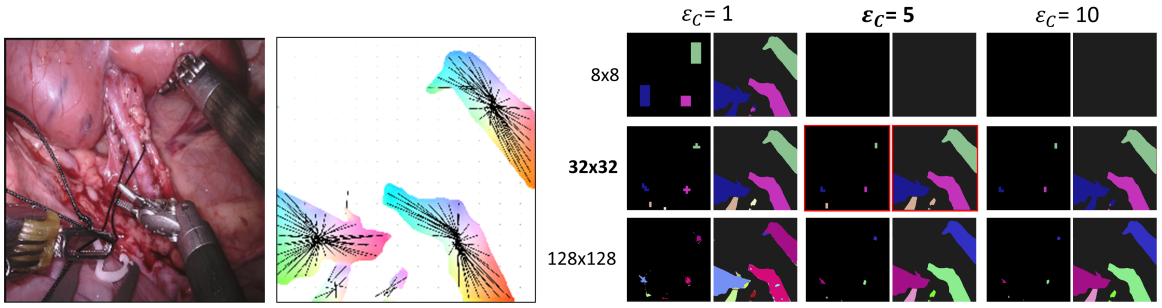

Figure 4.9: From left to right, original image, predicted displacement field, and examples of centroid regions and instantiation masks for different combinations of grid resolution and threshold ε_C. Mask colors indicate the ID assigned to the tube the instance belongs to. The combination adopted in our main experiments is highlighted in red.

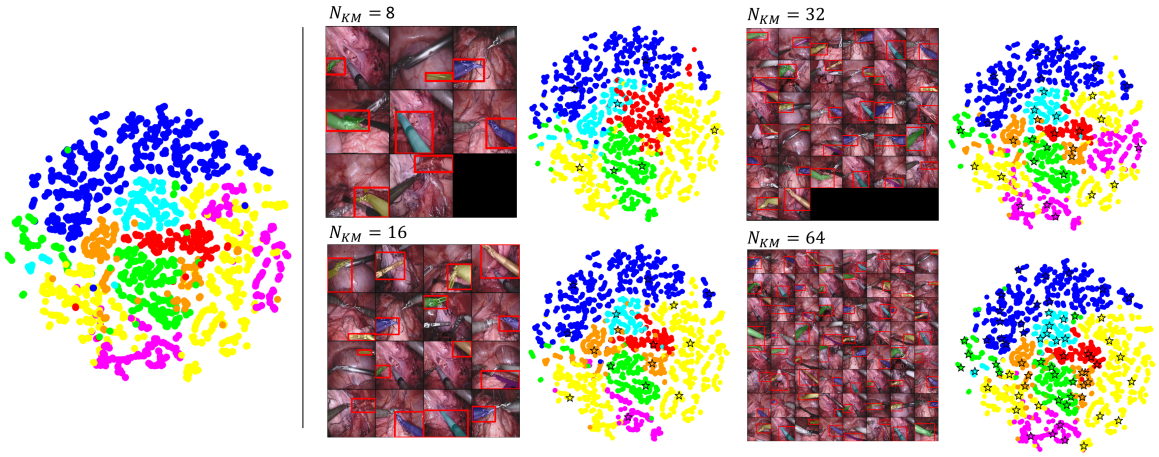

Figure 4.10: Left: visualization of the learnt feature representations of the EndoVis 2017 training set instances, projected in the 2D space using t-SNE algorithm [VdMH08]. Each instance point is colored according to the corresponding ground truth tool class. Right: K-Means++ clustering and prototype labels using different number of clusters N_{km}; projected features and prototype instances are colored accordingly to the corresponding prototype labels.

N_{km} used for K-Means clustering, equal to the number of prototype labels assigned by a potential human operator. In order to provide a complete overview, we present segmentation results obtained via instance classification by direct K-Means inference (KM), Teacher network prediction (T) and Student network prediction when trained using *sequence-wise* binary tool presence labels (S_S) and *frame-wise* (S_F). In addition, Figure 4.10 provides a visualization of the learnt feature distribution, the clustering process and the automatically selected prototype instances.

Result analysis provides different insights into the method. First of all, although a marginal improvement exists, increasing the number of prototype instances does not provide substantial performance gains for the Student network. This result may indicate that effective feature learning is a crucial methodological bottleneck, which cannot be solved by simply increasing the number of human-assigned labels. Secondly, the presented results highlight the consistent improvement in performance provided by the Student network, trained on weak labels matched through the Teacher model. Although the Teacher learns to substantially replicate K-Means clustering classification, as shown

N_{km}	Model			
	KM	T	S_S	S_F
8	43.86	45.66	52.64	**53.73**
16	38.52	41.34	50.02	52.64
32	42.33	45.26	51.23	52.44
64	**44.76**	**48.38**	**52.84**	53.37

(a)

N_{km}	Model			
	KM	T	S_S	S_F
8	30.47	30.88	42.21	45.86
16	**37.86**	**41.81**	**46.95**	47.33
32	32.49	36.40	46.40	**48.00**
64	36.10	41.02	46.91	47.96

(b)

N_{km}	Model			
	KM	T	S_S	S_F
8	**56.62**	56.80	**63.57**	63.38
16	56.03	57.25	60.63	61.96
32	56.11	**57.48**	62.80	**64.76**
64	53.80	57.22	62.24	63.88

(c)

N_{km}	Model			
	KM	T	S_S	S_F
8	54.08	54.14	57.75	58.03
16	**56.53**	**57.04**	57.40	58.53
32	55.53	55.86	**58.45**	59.48
64	55.52	56.02	57.92	**59.85**

(d)

Table 4.5: Results of the ablation study investigating the impact of the number of clusters N_{km} on final segmentation results, evaluated using challenge IoU metric. Results obtained using a): manually annotated binary masks on the EndoVis2017 dataset, b): FUN-SIS predicted binary masks on the EndoVis2017 dataset, c): manually annotated binary masks on the EndoVis2018 dataset, d): FUN-SIS predicted binary masks on the EndoVis2018 dataset. Best result across the number of clusters highlighted in bold.

by their similar performance, this is enough to perform a good weak label matching, responsible for Student's superior performance. Finally, a comparison between *frame-wise* (S_F) and *sequence-wise* (S_S) binary tool presence labels training, shows the value of using the latter, much cheaper, source of information, with the average gap between the two of $\sim \Delta 1.1\%$, consistently across datasets and binary mask sources.

4.6 Discussion

The results presented in Sections 4.4 & 4.5 prove the soundness of the proposed PAF-IS framework for instance segmentation. Our solution trains on endoscopic videos paired with binary segmentation masks, potentially obtained in an unsupervised way, and can incorporate weak information like binary tool presence labels. Human annotation effort is here limited to labelling a tiny set of prototype instances, automatically selected by our approach, with inexpensive classification labels: the ablation study presented in Section 4.5 shows that the size of such set can be reduced to 8 instances (~0.26% of the total number of training instances), with no significant performance drop. This result goes significantly beyond existing semi-supervised solutions like [ZJG+20], where a significant set of frames (up to 30%) needs to be labelled with pixel-wise annotations, while still providing inferior segmentation performance. Indeed, our complete pixel-wise annotation-free solution, using FUN-SIS predicted masks, outperforms fully-supervised and semi-supervised semantic segmentation approaches like MF-TN and DMF-TN by a consistent margin on the EndoVis 2017 dataset.
Although a performance gap still exists with top-performing fully-supervised instance segmentation approaches, we believe there exist several directions of improvement

to close such gap. First of all, temporal modelling could be easily learnt from the already available tracking information, currently exploited only at training time for feature learning. Secondly, as highlighted by the ablation study on clusters number, feature learning represents a crucial methodological bottleneck: if the learnt feature representations are sub-optimal, the unsupervised clustering may fail to separate tools belonging to different classes, hindering the following classifier training. In the current implementation, feature learning is performed in a completely unsupervised way, with no help from external information. Weak information about binary tool presence may be included at this stage to perform a more informed positive and negative feature sampling.

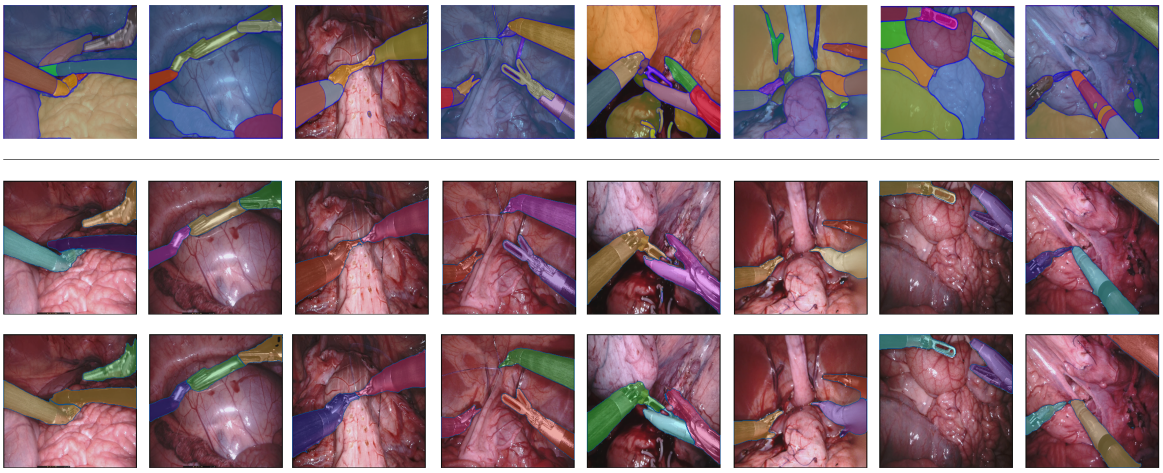

Figure 4.11: Top row: SAM [KMR+23] segmentation results on the EndoVis2017 dataset. Central row: PAF-IS instantiation results obtained from binary annotated masks only. Bottom row: PAF-IS instantiation results obtained from FUN-SIS predicted binary masks only.

In addition to these direct improvements, PAF-IS, not requiring pixel-wise semantic labels, can leverage recent break-through solutions like SAM [KMR+23] (Segment Anything Model) to directly obtain instance-wise masks for following feature learning and instance-wise tool type classification. Figure 4.11 shows qualitative results from SAM (without text prompts, not yet released as of now, April 2023) on the EndoVis2017 dataset, compared to PAF-IS predictions. Even if SAM segmentation results are currently over-segmenting tools, breaking them up into individual parts, our PAF-IS instantiation predictions could be used to group those parts, exploiting the high-quality boundary segmentation that SAM can already provide.
In conclusion, PAF-IS major contribution relies on its ability to lift the need for pixel-wise semantic and instance annotations of the training data. This may open up new research directions aimed at better exploiting human annotation effort, for example by focusing it on particularly representative or challenging samples.

4.7 Conclusion

Overall, Chapters 3 & 4 thoroughly explored the use of different sources of weak knowledge about surgical tools, easily obtainable and highly repurposable across surgical domains, to solve the problems of binary and instance tool segmentation. Their integra-

tion de-facto removes the need for pixel-wise manual annotations to solve the instance segmentation task, yielding results superior to the ones of several state-of-the-art fully-supervised approaches. While these two contributions show that weak prior knowledge can be sufficient to address the tool localisation and identification problem in the image space, the next Chapter extends the scope of instrument localisation to the 3D space. Strong prior knowledge, in the form of 3D kinematic modelling, is integrated into a Deep Learning architecture, allowing to solve the 3D pose estimation task without relying on manual annotations.

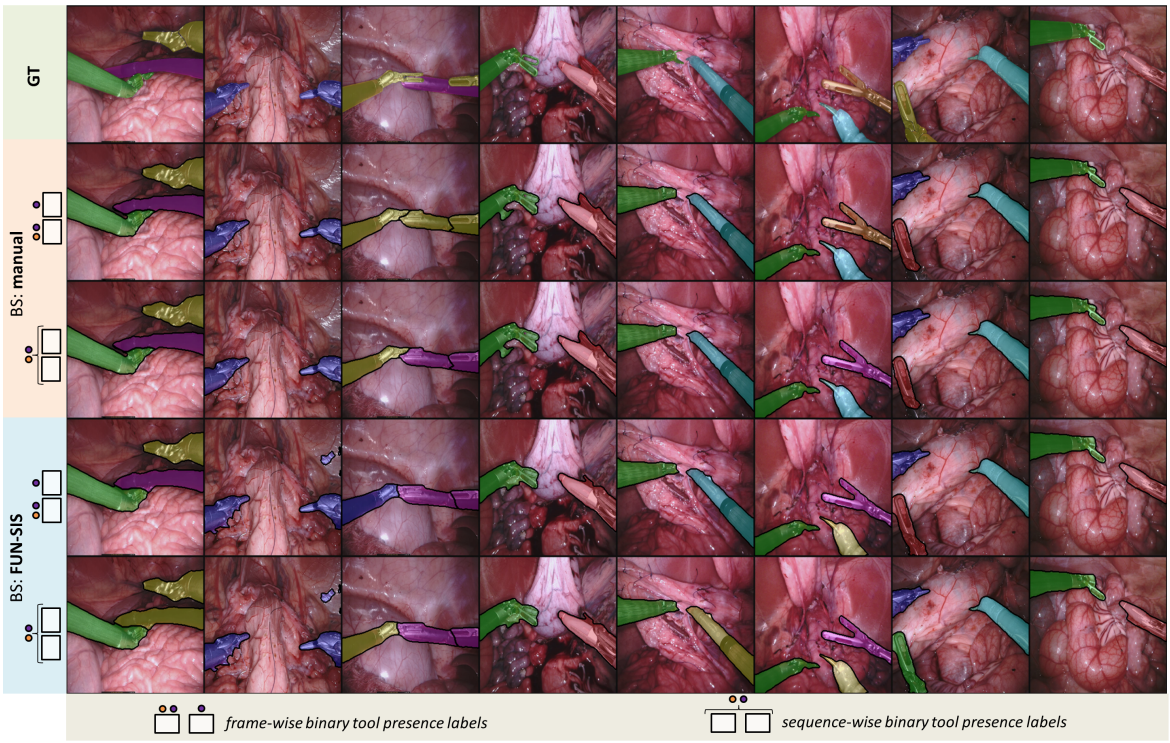

Figure 4.12: Qualitative segmentation results from the EndoVis2017 dataset. Row 1: ground truth; rows 2-5: PAF-IS Student trained on (2) manually annotated binary masks and frame-wise tool presence labels, (3) manually annotated binary masks and sequence-wise tool presence labels, (4) FUN-SIS predicted binary masks and frame-wise tool presence labels, (3) FUN-SIS predicted binary masks and sequence-wise tool presence labels.

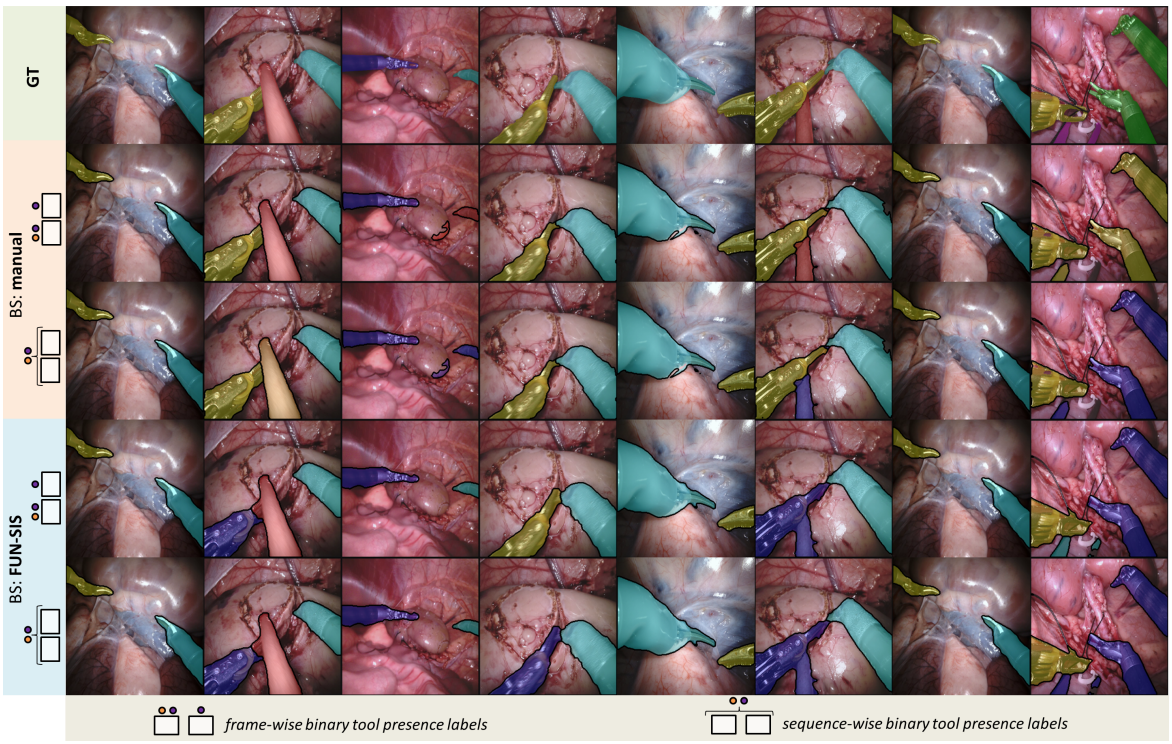

Figure 4.13: Qualitative segmentation results from the EndoVis2018 dataset. Ground truth and PAF-IS Student results are presented in the same order as Figure 4.12 above.

CHAPTER 5

KI-BOT: a Kinematic Bottleneck Approach For Pose Regression of Flexible Surgical Instruments directly from Images

Contents

5.1 Introduction 118
5.1.1 Objective & Contributions 119
5.2 Methodology 119
5.2.1 Method Overview 120
5.2.2 Backgroundizer module: an Inpainting Problem 121
5.2.3 Camera and Robot Modelling 122
5.2.4 Regressor and Decoder 123
5.3 Experimental Set-up 123
5.3.1 Robotic System 123
5.3.2 Datasets 124
5.3.3 Design Choices & Training Details 126
5.4 Experiments and Results Analysis 126
5.4.1 Backgroundizer and Physical Modules Results 127
5.4.2 Kinematic Regression Results 127
5.5 Discussion 129
5.6 Conclusion 132

5.1 Introduction

In Chapters 3 and 4 we addressed the problem of image-level tool localisation exclusively relying on weak prior knowledge. In this Chapter we rethink the problem of tool localisation when stronger problem knowledge is available. The 3D pose of surgical instruments is an extremely valuable piece of information for robotic automation, enabling applications like dynamic motion constraints [MISF20] and visual-servoing [ZRCM+21]. As discussed in Section 1.4.1.2, parametrized 3D modelling of surgical tools offers the opportunity to tackle the 3D pose estimation problem from a pure vision-based standpoint. While the general problem formulation shown in Figure 5.1, top, has been adopted by several works, we identify three common critical aspects, that the work in this Chapter tries to address:

1. *inference-time optimization*: most of the existing solutions use the general framework shown in Figure 5.1, top, to iteratively refine measured kinematic values at inference time. Inference-time optimization often results in low throughput, incompatible with real-time needs;

2. *need for manual annotations*: in order to project the endoscopic image into the same space as the rendered image, existing approaches learn P_{real} from a set of labelled data, usually manually annotated for the tool segmentation task;

3. *validation domain*: validation of existing approaches has been mostly carried out for rigid endoscopic tools. For such tools the reliability of the recorded kinematics is usually superior compared to the one of flexible endoscopic tools, more subjected to deformation due to tool-tissue interactions.

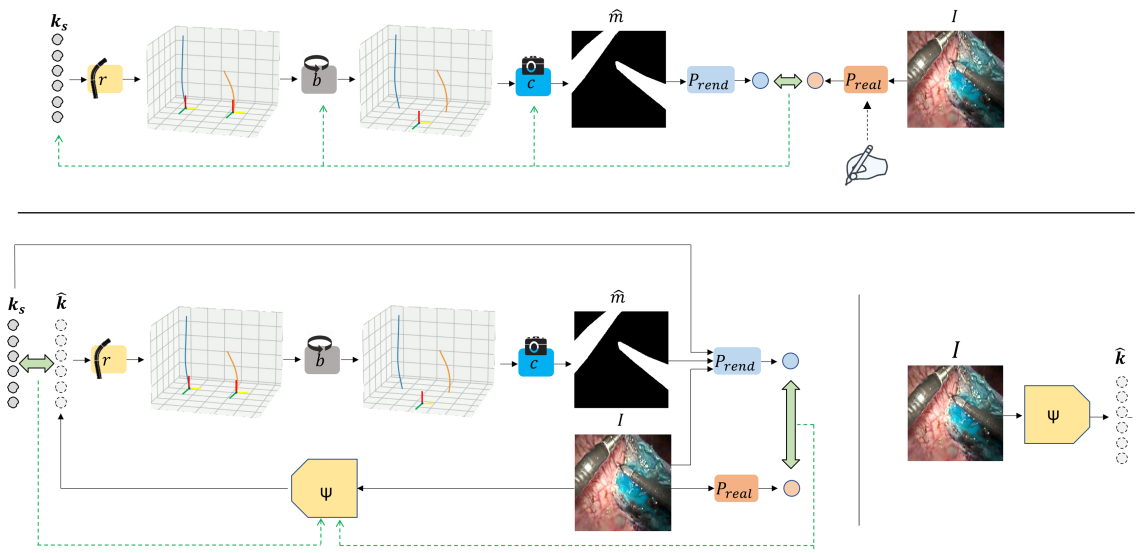

Figure 5.1: Application of the general framework for vision-based 3D pose estimation presented in Chapter 1, Figure 1.12. Top: existing approaches commonly learn P_{real} from annotated datasets, and use the distance between the two projections P_{real} and P_{rend} as optimization error to refine the kinematic values $\mathbf{k_s}$, recorded by the robotic system. Bottom left: our solution shifts the target of the optimization to the neural network model Ψ, which directly regresses the kinematic vector $\hat{\mathbf{k}}$ from images. Bottom right: at inference time the predicted kinematics $\hat{\mathbf{k}}$ is obtained from the input image only, by forward propagation through Ψ.

5.1.1 Objective & Contributions

In this Chapter we present KI-BOT, a KInematic BOTtleneck approach for 3D pose estimation of flexible surgical instruments directly from images. In order to address the above-mentioned criticalities, we propose a fully-differentiable architecture, training a deep learning model to directly estimate kinematic values from images, without requiring manual annotation of the training data. Compared to existing solutions, KI-BOT (Figure 5.1, bottom):

- trains a regressor model to directly estimate, from single frames, the 3D pose of the instruments, parametrized by means of their kinematic joint values. Optimization is performed at training time only, in order to learn the parameters of the regressor model. This makes kinematic estimation at inference time a single forward propagation through the trained regressor;
- the full architecture trains end-to-end by exploiting an auto-encoder formulation bottlenecked by the presence of the physical model of instruments and endoscopic camera. Such architecture is designed to avoid the use of manual annotations for P_{real} estimation, by leveraging the weak complementary information provided by the recorded kinematics;
- is validated using a flexible endoscopic robot, in multiple datasets including one containing challenging in-vivo endoscopic submucosal dissection procedures.

5.2 Methodology

The 3D pose estimation problem is here formalized as regressing from an endoscopic image the value of the n kinematic joints $\mathbf{k} = \{k_0, k_1, ..., k_{n-1}\}$ describing the configuration of the robotic system instruments in such frame. The kinematic joint values, combined with a 3D kinematic model of the instruments, allow to reconstruct their complete 3D shape. In practice, the problem is here formulated as training the regressor model $\psi : I \rightarrow \hat{\mathbf{k}}$ that maps an image I, containing the instruments in a configuration represented by the set of ground truth joint values $\mathbf{k} = \{k_0, k_1, ..., k_{n-1}\}$. If $\hat{\mathbf{k}}$ is known, the training problem can be formulated as minimizing the fully-supervised loss L_r^{FS} between $\hat{\mathbf{k}}$ and the estimated joint values $\hat{\mathbf{k}}$, regressed by ψ from I:

$$L_r^{FS} = |\hat{\mathbf{k}} - \mathbf{k}|. \tag{5.1}$$

However, $\mathbf{k}$ is not known in practice. The measured joint values $\mathbf{k_s}$, recorded by the robotic system, are generally inaccurate, due to tool-tissue interactions and possible unmodelled non-linearities. Therefore $\mathbf{k_s}$ cannot be directly used as supervision signal to train ψ.
The general framework shown in Figure 5.1, top, adopted by several state-of-the-art approaches, introduces in the problem the 3D kinematic model of the instruments and the model of the endoscopic camera. This allows to map the kinematic values to the corresponding tool projections in the image space $\hat{m}$. The problem is then commonly solved by iteratively refining the kinematic joint values $\mathbf{k_s}$, minimizing the distance between the two projections $P_{real}(I)$ and $P_{rend}(\hat{m})$. However, as previously discussed, this open-loop formulation normally requires the use of manual annotations to learn P_{real}.

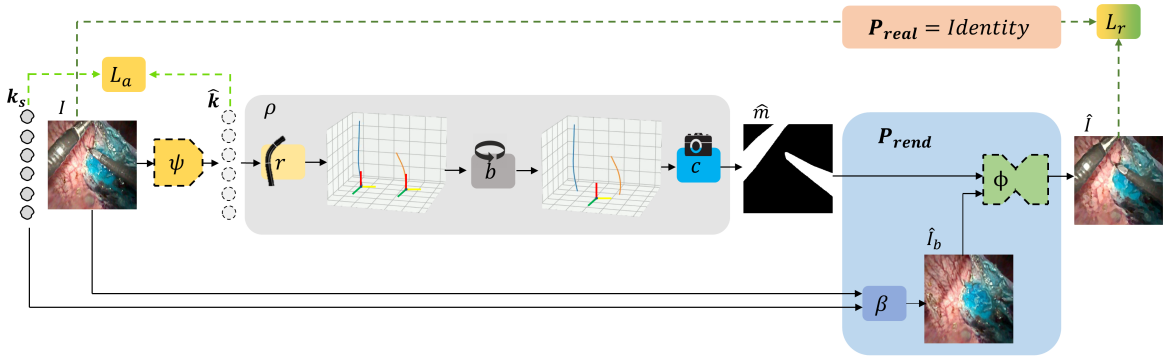

Figure 5.2: Full KI-BOT training architecture: given an input image I and the corresponding measured kinematics $\mathbf{k_s}$, the training architecture forces a separation between the appearance of the image and its kinematic content. In the bottom branch, the module β transforms I into a backgroundized version of itself $\hat{I}_b$. In the top branch, the regressor ψ reduces I to the estimated kinematic configuration $\hat{\mathbf{k}}$. $\hat{\mathbf{k}}$ is then mapped into the binary silhouette projection $\hat{m}$ of the instruments on the image plane by means of the physical model ρ, consisting of the 3D model the instruments r, the robot-camera transformation b and the model of the endoscopic camera c. A decoder ϕ tries to reconstruct the input image I from $\hat{I}_b$ and $\hat{m}$. The model is trained by means of the image-based loss L_r, helped by the auxiliary loss L_a. The backgroundizer β and the physical module ρ are trained in advance and frozen during the training of regressor and decoder.

5.2.1 Method Overview

In order to avoid relying on manual annotations, we propose a training architecture built as an auto-encoder, trained by reconstructing the input image I as output of a decoder model ϕ. Figure 5.1, bottom left, provides an overview of our approach, highlighting how it differs from the commonly adopted framework (Figure 5.1, top). Figure 5.2 details our architecture, showing how it is effectively implemented.
Our solution aims at separating the appearance of the image from its kinematic content. In order to do that, the architecture presents two separate branches:

- the *appearance* branch consists of a module β, called *backgroundizer module*, which converts the input image I to the approximate background-only version of itself $\hat{I}_b$;
- the *content* branch, which contains the model ψ, regressing kinematics from the image.

However, a general auto-encoder formulation featuring these elements only would fail to effectively train the regressor ψ: without any constraints the vector $\hat{\mathbf{k}}$, regressed by ψ, would not necessarily contain only kinematic information, and it would not necessarily be physically meaningful (i.e: each regressed value corresponding to a specific instrument joint). For this reason, the regressor ψ is followed by the *physical* module ρ, a model of robotic instruments and endoscopic camera. The module maps the estimated kinematics $\hat{\mathbf{k}}$ to the instrument 3D shape, through a forward kinematic model of the instruments, and reprojects it to the camera plane as a binary tool mask $\hat{m}$. The *physical* module ρ takes the pivotal role of a *kinematic bottleneck* in the architecture: because of the way it processes the low dimensional vector $\hat{\mathbf{k}}$, it gives it the explicit meaning of *kinematics*, forcing, in turn, ψ to learn the expected regression transformation.

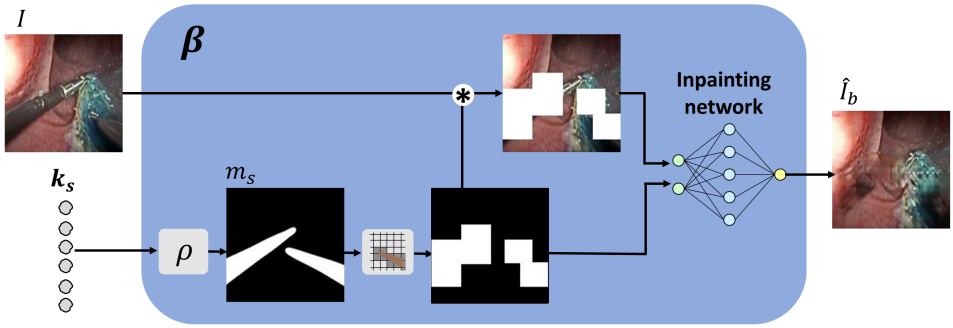

Figure 5.3: Backgroundizer module β: the imprecise kinematics $\mathbf{k_s}$ is converted to the binary projection representation m_s, through a robot-renderer model equivalent to ρ. m_s is then expanded to account for uncertainties and used to mask the image I associated with $\mathbf{k_s}$, which is then fed to an inpainting network which produces $\hat{I}_b$, the backgroundized version of I.

The outputs of the two branches, $\hat{I}_b$ and $\hat{m}$, are then fed to the decoder ϕ, which reconstructs the image $\hat{I}$ from the combination of the two. The loss L_r, measuring the distance between I and $\hat{I}$ is minimized to train the regressor model ψ.

The proposed architecture can be seen as a modification of the general framework for vision-based 3D pose estimation, with P_{real} being an identity mapping, and P_{rend} being an operation mapping the tool projection $\hat{m}$ to the reconstructed image $\hat{I}$, learnt with an unsupervised formulation as discussed in the next paragraph.
The *backgroundizer* and the *physical* module are discussed in sections 5.2.2 and 5.2.3.

5.2.2 Backgroundizer module: an Inpainting Problem

In order to make the reconstruction task feasible, the decoder ϕ must be provided with information about the appearance of the image. [JGBV20], adopting a similar architecture for the task of human pose-estimation, provide this information by feeding a second image I' to the decoder, shifted in time with respect to I (with I, I' belonging to the same video). This approach is effective as long as I' contains the same exact background as I and the subject in it has a different pose than the one the model is trying to regress. Unfortunately, none of the hypotheses can be verified in the complex surgical scenario: robotic instruments can remain in the same configuration for relatively long periods during surgery, and the background appearance is constantly modified by the direct and indirect action of the instruments (pulling tissue, bleeding, smoke etc.) and by the movement of the camera. To address this issue, we introduce a novel *backgroundizer module* β that informs the decoder about the background of the image I, by estimating the *backgroundized* image $\hat{I}_b$ as:

$$\hat{I}_b = \beta(I, \mathbf{k_s}). \tag{5.2}$$

The image $\hat{I}_b$ is obtained by exploiting the rough instrument localisation provided by the recorded kinematics $\mathbf{k_s}$, accounting for its uncertainty. The full pipeline for the *backgroundizer module* can be observed in Figure 5.3. In order to leverage the imprecise information provided by $\mathbf{k_s}$, we first map it to the corresponding binary mask m_s through a *physical* module, equivalent to ρ. We then take into account $\mathbf{k_s}$ uncertainty by expanding the binary mask m_s, in order to cover a greater area in the image space. To do that, we simply overlap a $a \times b$ grid on the mask m_s, and assign to a whole grid quadrant a

value of 1 if at least one pixel of m_s inside it corresponds to a tool. The masked image obtained by multiplying the image I by the expanded mask is then processed by an inpainting network, to obtain the *backgroundized* image $\hat{I}_b$: the problem is formulated as image inpainting from block occlusion [EAAMA19], solved using the inpainting network implementation proposed in [LRS+18]. The inpainting network can be trained on artificial data, obtained by collecting background-only images, randomly masking them and training the network to reconstruct the original images.

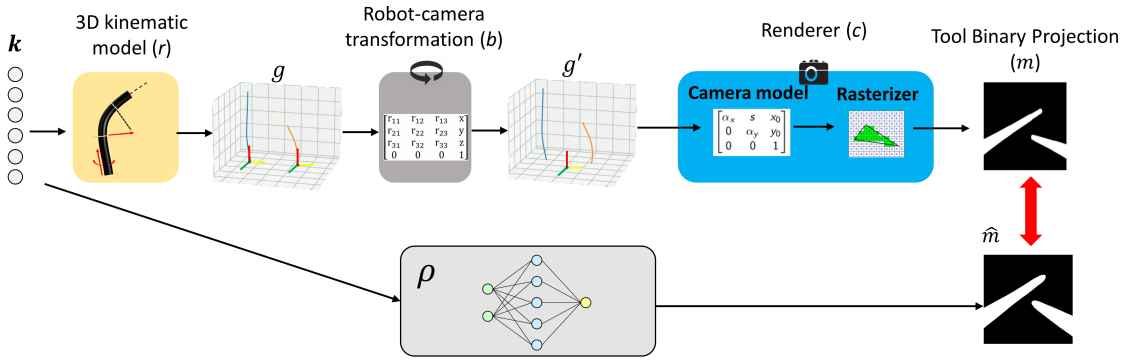

Figure 5.4: Physical module ρ training: a synthetic dataset is generated by mapping kinematic vectors $\mathbf{k}$ *to the corresponding projection masks m, using the 3D kinematic model of the tools, the robot-camera transformation matrix and a potentially non differentiable renderer. The mapping between kinematic vectors and projection masks is directly learnt by the neural network ρ, trained on the synthetic dataset.*

5.2.3 Camera and Robot Modelling

The *physical* module ρ, following the regressor ψ, has the crucial role of *kinematic bottleneck*, giving $\hat{\mathbf{k}}$ the explicit meaning of *kinematics*, and forcing ψ to properly learn the kinematic regression task. This module consists of a geometrical, forward-kinematic model of the instruments, followed by a renderer (pinhole model of the endoscopic camera and rasterizer). The forward kinematic model of an instrument is defined as the mapping $r : \mathbf{k} \rightarrow g$, with g being the 3D shape of the instruments, referred to the instrument reference frame. Assuming that the parameters of the transformation b between such reference frame and the robot-mounted camera reference frame are known from hand-eye calibration, and that the camera is calibrated, the instrument 3D shape g can be projected on the camera plane and rasterized in pixel coordinates. The full rendering operation is defined as $c : g \rightarrow \hat{m}$, with $\hat{m}$ being the binary tool segmentation mask, obtained binarizing the projection of the 3D shape of the instruments g on the camera plane.
Given their embedding in an end-to-end trainable neural network architecture, both the instrument model r and the renderer c are required to be differentiable, in order to allow the error signal to be backpropagated through them. In this work, we address these challenges by directly learning the mapping between the kinematic vector $\hat{\mathbf{k}}$ and the output binary mask $\hat{m}$ (Figure 5.4). Given the forward kinematic model of the instruments and a renderer (both not necessarily differentiable), a virtually infinite set of coupled samples $\{\mathbf{k}, m\}$ can be generated and used to train a neural network to learn the direct mapping $\rho = c \odot b \odot r : \hat{\mathbf{k}} \rightarrow \hat{m}$. The neural network is implemented as a GAN generator, mapping

the input kinematic vector $\mathbf{k}$ to the corresponding tool projection mask $\hat{m}$. The training loss is formulated as a standard binary cross-entropy loss between $\hat{m}$ and m.

5.2.4 Regressor and Decoder

The regressor ψ processes the input image I to predict the corresponding instrument kinematic configuration $\hat{\mathbf{k}}$. Given the general independence between instruments (if more than one is simultaneously present), ψ is implemented using a shared backbone and l separate heads, one per robotic instrument of the robotic system. The backbone is based on ResNet-50 [BCF18], using only the first and second convolutional layers. Each head consists of four convolutional layers, followed by global average pooling and a three-layer fully connected network, having n/l units in output. In order to take advantage of the rough information provided by the measured kinematics $\mathbf{k_s}$, without using it as a strong supervision signal, we introduce a *soft mean squared error* auxiliary loss function L_a, defined as:

$$L_a = \frac{1}{n}\sum_{i=1}^{n} max((k_i - \hat{k}_i)^2 - t^2, 0), \tag{5.3}$$

with t being a *tolerance* hyperparameter that can be interpreted as a maximum accepted offset of $\hat{\mathbf{k}}$ with respect to $\mathbf{k_s}$. The idea behind this loss is to ease the optimization process by providing the regressor ψ with a range of kinematic values in which the solution is likely to be found, avoiding any hard-coded constraint.

The decoder ϕ processes the predicted projection mask $\hat{m}$ and the *backgroundized* image $\hat{I}_b$ to reconstruct the image $\hat{I}$. The network is implemented as a *UU-net*, an extension of the well-established *U-net* architecture [RFB15], having two separate contracting paths for $\hat{m}$ and $\hat{I}_b$, and an expanding path where corresponding features from the two contracting branches are concatenated. The reconstructed image $\hat{I}$ is compared to the input image I by means of a perceptual loss L_r, defined as:

$$L_r = \|\Gamma(I) - \Gamma(\hat{I})\|^2, \tag{5.4}$$

where Γ is a feature extractor implemented as a VGG-16 network [SZ14] pre-trained on ImageNet dataset [DDS+09]. Regressor ψ and decoder ϕ are jointly optimized to minimize L_r.

5.3 Experimental Set-up

5.3.1 Robotic System

The experimental validation was performed using the STRAS robot [DDZZ+13], a teleoperated prototype for flexible robotic endoscopic surgery [LDH+17]. The robot is built as a standard endoscope having two operating instrument channels, through which robotic arms can be positioned. The robot arms used in STRAS are flexible cable-actuated instruments. Each instrument has 3 joint angles/positions, resulting in a total dimension of $\hat{\mathbf{k}}$ equal to 6: rotation around the instrument's main axis, translation along the same axis, and a cable-actuated bending, defined by the delta between the lengths of the two cables used for actuation (Figure 5.5).

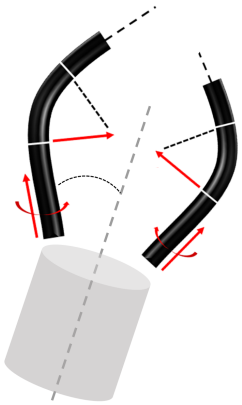

Figure 5.5: STRAS robotic system model, showing the two instruments, each one having 3 joint angles/positions: rotation around the instrument main axis, translation along the same axis, and a cable-actuated bending, here represented as the radius of curvature, but actually measured as the delta between the lengths of the two cables used for actuation. Instrument's main axis forms a small angle (~ 10°) with the main axis of the endoscope (grey cylinder).

The instrument main axes are parallel to the exit parts of the instrument channels, which deviate from the endoscope main axis, forming a small angle (~ 10°), resulting almost perpendicular to the camera plane. During working configuration, instruments are bent, in order to be inside the field-of-view of the camera; therefore, when the rotational joint is activated, the visible part of the instruments moves on a plane almost parallel to the camera plane, avoiding ambiguities between pose and projected shape [AOH+18]. Using a constant curvature assumption, the joint angles can be used to compute the forward kinematic model of the robotic instruments, following equations detailed in [WIJ10]. The robot instruments are teleoperated by the user sending commands through a master console. The positions of the motors are recorded into the array of joint values $\mathbf{k_s}$. Camera images are acquired through an acquisition board in a synchronized fashion, resulting in RGB images with 570 × 760 pixel resolution. The forward kinematic model of the instruments and the renderer were implemented using the VTK library [GSBW12]. The two VTK models were used exclusively in the datasets generation process, as detailed later in this paragraph, and not included in the training architecture, given their non-differentiability.

5.3.2 Datasets

Three different datasets were used for validating the proposed approach, two from real acquisitions and a semi-synthetic one:

- ***phantom*** dataset: this dataset was built on the bench-top, using a plastic phantom model of the human digestive system. During the acquisition, both the endoscope and the plastic model were moved, in order to avoid having a static background appearance.
- ***in-vivo*** dataset: videos for this dataset were collected as part of a pre-clinical study [MLF+19] from a 4-day set of *in-vivo* experiments on porcine models[1]. This

[1]The study protocol for this experiment was approved by the Institutional Ethical Committee on Animal

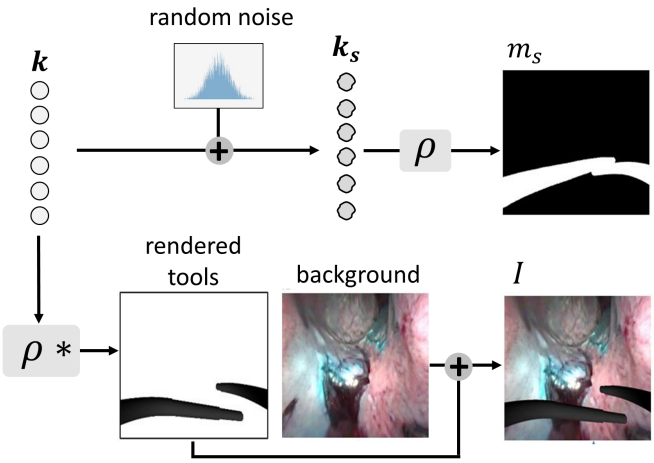

Figure 5.6: Semi-synthetic dataset building: a kinematic configuration **k** *and a background b are randomly sampled.* **k** *is then fed to a VTK model of the robotic instruments (ρ∗, where * means that binarization of the rendered image is not applied, differently from ρ) and rendered on the image space. The projection is then blended with b to obtain the image I. Parallelly, random noise is added to* **k** *to simulate the measured kinematics* $\mathbf{k_s}$ *(visualized as the corresponding binary projection* m_s *on the image plane obtained through ρ, in order to qualitatively show the difference with the GT kinematic configuration* **k***).*

dataset presents several challenges compared to standard benchmark datasets like EndoVis, including low foreground/background contrast, highly cluttered and changing background, frequent occlusions and bleeding, and strong tool-tissue interactions.

- ***semi-synthetic*** dataset: this dataset was created by blending background-only images, automatically extracted from the *in-vivo* dataset by parsing the associated kinematic information (although imprecise), with rendered robot instruments, obtained using the VTK model of tools and renderer, and random kinematic configurations **k**. In order to simulate a realistic imprecise kinematic information $\mathbf{k_s}$ for this dataset, we added to the nominal value **k** a normally distributed noise, whose range was empirically chosen to match the real-dataset noise. The full process followed to build the semi-synthetic dataset is shown in Figure 5.6.

For the *semi-synthetic* dataset, the ground truth (GT) joint configuration is known a-priori. For the *phantom* and *invivo* datasets, the GT joint configuration is unknown, unless external sensors are introduced. The evaluation on these two datasets is therefore performed indirectly: the 3D shape of the instruments is reconstructed according to the predicted kinematics $\hat{\mathbf{k}}$, and projected on the camera plane: the projected mask $\hat{m}$ can be then compared to the instruments ground truth location on the images, obtained via manual segmentation, providing an indirect evaluation of the estimated $\hat{\mathbf{k}}$.

In order to train and evaluate the *physical* module, a synthetic dataset was built generating random couples $\{\mathbf{k_i}, m_i\}$, using the VTK model of instruments and renderer.
For the training of the inpainting network, two separate training datasets were artificially built. The first one, *real backgrounds*, was built automatically extracting background-

Experimentation (ICOMETH No.38.2011.01.018). Animals were managed in accordance with French laws for animal use and care as well as with the European Community Council directive no. 2010/63/EU

Dataset	# Images Training	# GT testing images
physical module	100k	10k
real backgrounds	6000	/
phantom backgrounds	1 (+augmentation)	/
semi-synthetic	20400	2400
phantom	6800	800*
in-vivo	28800 (4 days)	400*/day

Table 5.1: Number of images in each training and testing dataset (GT 2-D segmentation mask only, obtained via manual segmentation)*

only images from the *in-vivo* dataset, according to the associated kinematic information (although imprecise), and used to train the *backgroundizer* for the *semi-synthetic* and *in-vivo* experiments; the second one, *phantom backgrounds* was built by extracting a single background-only image from the *phantom dataset*, strongly augmented through operations like rotation, cropping, lighting etc., and used for the *phantom* experiments. Table 5.1 summarizes the number of images and GT images for all the datasets.

5.3.3 Design Choices & Training Details

As a preliminary step, the inpainting network, belonging to the *backgroundizer module* β, and the neural network implementation of the *physical* module ρ were trained. The *physical* module was trained on the randomly generated couples $\{\mathbf{k_i}, m_i\}$, for 100 epochs, using a batch size of 64 and a learning rate of 0.0005. For the inpainting network the grid resolution was set to 6×8, as experimentally found providing the best trade-off between area covered and inpainting quality. Two inpainting networks were trained (on *real backgrounds* and *phantom backgrounds*), with a batch size equal to 16 and a learning rate of 0.001, until visually satisfying results were reached.
The two modules were then frozen during the end-to-end training of the regressor ψ and the decoder ϕ. The regressor and the decoder were trained alternatively (1 iteration each) using the Adam optimizer with learning rates equal to 1e-4 and 1e-5, respectively, a batch size of 36, for 100 epochs: the decoder was trained on the loss L_r; the regressor was trained on a weighted sum of losses L_r and L_a, defined as:

$$L = \alpha_a L_a + \alpha_r L_r, \tag{5.5}$$

with α_a, α_r being weight parameters set to 10 and 0.001, respectively, in order to balance the magnitude of the two losses. The tolerance parameter t for each joint was empirically determined in the context of other work using the STRAS robotic system [dCRPR19, DDZZ+13].

5.4 Experiments and Results Analysis

In this Section we present the validation results obtained by the proposed approach on the three datasets.
First, in Section 5.4.1, we present results for the *backgroundizer* β and the *physical* module ρ.

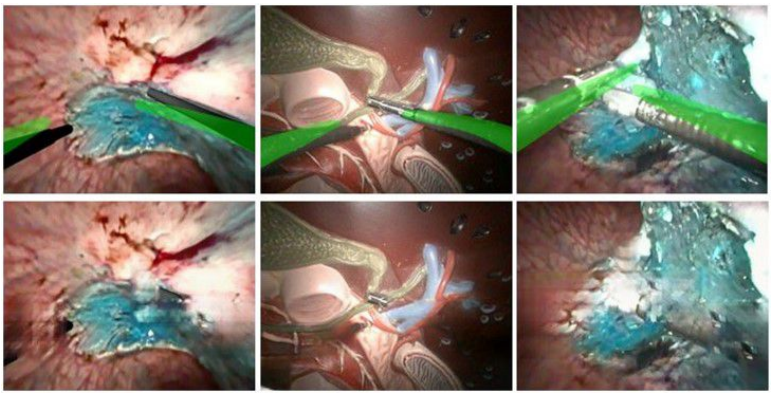

Figure 5.7: Examples of backgroundizer module results. Top row shows one image for each dataset (semi-synthetic,phantom, in-vivo), with the corresponding measured kinematics $\mathbf{k_s}$ *projection (imprecise). Bottom row shows the corresponding backgroundized versions, obtained using the rough localisation provided by* $\mathbf{k_s}$.

Then, is Section 5.4.1, we present and analyze results for kinematic estimation. In order to assess the actual contribution of the method, we compare its performance against two baseline models. They both consist of solely the regressor model ψ trained by minimizing only the loss function L_a defined in eq. 5.3:

- the first baseline model $BSupK_s$ was trained by setting the tolerance parameter t to 0, thus resulting in a fully-supervised training, having $\mathbf{k_s}$ as target;
- the second baseline model $BSupSoftK_s$ was trained by setting the parameter $\mathbf{t}$ to the same value as the one used for our main model, thus resulting in a softer supervision.

Together with the baselines, also the raw measured kinematics $\mathbf{k_s}$ provided by the robotic system is evaluated.
For the evaluation on the *in-vivo* dataset, leave-one-out cross-validation was performed, by training the models on data from 3 days and testing on the remaining. The average of the results across the 4 days was then computed and reported.

5.4.1 Backgroundizer and Physical Modules Results

The *physical* module ρ was evaluated by means of the Intersection-over-Union metric, resulting in an IoU of 94.3% on the testing dataset.
For the *backgroundizer* module, the inpainting network was qualitatively evaluated. Qualitative results on the three datasets are shown in Figure 5.7.

5.4.2 Kinematic Regression Results

For the *semi-synthetic* dataset, the GT kinematic configuration $\mathbf{k}$ of the instruments is known, and the Mean Absolute Error (MAE) for each joint can be computed. Reprojection of the estimated instrument shape on the image plane is also using the IoU metric. Results are reported in Table 5.2. For the real *phantom* and *in-vivo* datasets the reprojection error with respect to the manually annotated GT is evaluated, via IoU metric. Results are

	left tool				right tool			
	tr.[mm]	ro.[deg]	be.[mm]	IoU[%]	tr.[mm]	ro.[deg]	be.[mm]	IoU[%]
$\mathbf{k_s}$	6.40	16.40	0.87	24.41	3.21	13.07	0.77	42.21
$BSupK_s$	4.00	10.30	0.73	64.62	1.85	7.05	0.49	64.68
$BSupSoftK_s$	4.27	11.00	0.62	55.56	2.13	6.97	0.50	62.03
KI-BOT	**1.75**	**6.02**	**0.47**	**73.86**	**1.17**	**3.61**	**0.30**	**85.42**

Table 5.2: Semi-synthetic dataset results. Comparison with raw kinematics $\mathbf{k_s}$ *and fully supervised methods* $BSupK_s$, $BSupSoftK_s$. *For the joint mean absolute errors (translation: tr., rotation: ro., bending: be.), lower is better. For the reprojection IoU metric higher is better.*

	left tool		right tool	
	phant.	*in-vivo*	*phant.*	*in-vivo*
$\mathbf{k_s}$	28.32	42.14	32.73	44.73
$BSupK_s$	28.02	48.21	32.72	46.12
$BSupSoftK_s$	31.31	45.83	33.02	43.64
KI-BOT	**64.00**	**55.40**	**72.52**	**55.43**

Table 5.3: Evaluation of the IoU on the real datasets (Phantom & in-vivo). Comparison with raw kinematics $\mathbf{k_s}$ *and fully supervised methods* $BSupK_s$, $BSupSoftK_s$. *For the in-vivo the average of the results obtained for each day is reported.*

reported in Table 5.3. Qualitative results for the three datasets can be observed in Figure 5.8.

The results obtained in the three datasets confirm the effectiveness of our solution to learn kinematic regression from images without relying on manual annotations. This is shown directly by the improved kinematics estimation in the *semi-synthetic* dataset, and indirectly by the higher reprojection accuracy on the *semi-synthetic* and real datasets, with respect to both the baselines and the raw kinematics.

In the *semi-synthetic* dataset KI-BOT improves the accuracy of each joint value compared to the recorded kinematics $\mathbf{k_s}$: -3.345 mm MAE (-68.11% error) for translation, -9.92 deg MAE (-67.83% error) for rotation and -0.435 mm MAE (-53.51% error) for bending, on average between left and right tool. This results in a largely improved IoU between reprojected tools and GT tool masks (+Δ38.04% IoU). A similar improvement for the reprojection accuracy with respect to $\mathbf{k_s}$ is achieved in the *phantom* dataset (+Δ37.74% IoU). In the in-vivo dataset such improvement, while still consistent, is reduced: +Δ11.98% IoU. This can be explained by two main reasons: 1) the dataset is more challenging due to the factors mentioned in section 5.3 (e.g. low instruments-background contrast), which can affect image-based training; 2) phenomena such as tool-channel interaction and slackening [Cab16], more evident under strong tool-tissue interaction conditions.

When compared to the baselines models, KI-BOT outperforms both $BSupK_s$ and $BSupSoftK_s$ by consistent margins: +Δ14.96% IoU and +Δ20.85% IoU in the *semi-synthetic* dataset, +Δ37.89% IoU and +Δ36.06% IoU in the *phantom* dataset, +Δ8.25%

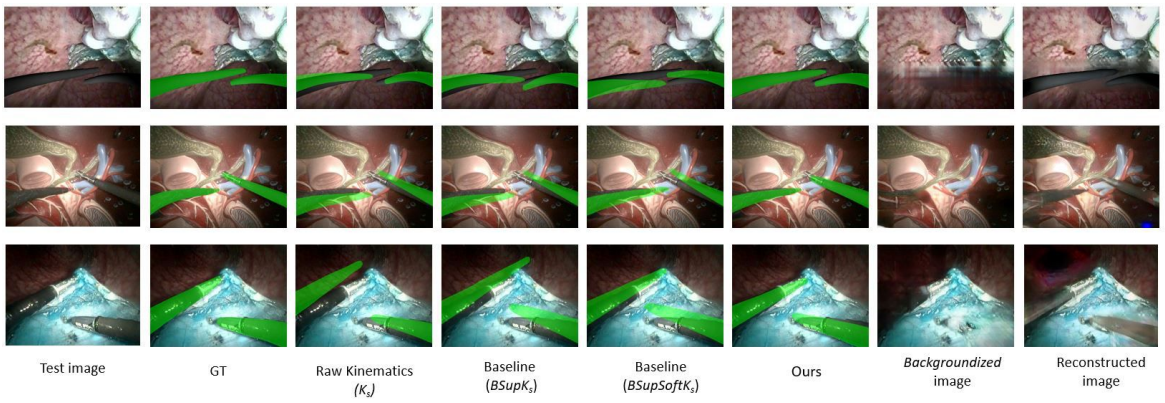

Figure 5.8: Examples of qualitative results on the three testing datasets. Top row: semi-synthetic; middle row: phantom; bottom row: in-vivo. The last two columns show corresponding reconstructed image and backgroundized image. Note that at inference time none of the two is needed/obtained, since the image-based regressor ψ is a completely independent module.

IoU and +Δ10.68% IoU in the *in-vivo* dataset (average for left and right tool). Interestingly, also the baseline models, directly supervised by $\mathbf{k}_s$ during training, improve the quality of kinematics, although by a smaller amount than KI-BOT. This can be seen as an empirical confirmation of the *unpredictability* property, discussed in Chapter 3: as the noise affecting $\mathbf{k_s}$ is not predictable from the image, the regressor learns the *easiest* pattern, by averaging out the noise across the training data.
The evaluation of $BSupSoftK_s$ also provides an ablation study of our method, showing that the improvements brought by KI-BOT are not solely caused by the soft loss, but mostly derive from the image-based loss L_r.

Finally, the average processing time for each image, without any specific model optimization, is approximately 30 ms (~ 33 fps) on a single Tesla V100 GPU, compatible with real-time needs.

5.5 Discussion

The proposed KI-BOT solution aimed at showing the feasibility of tackling the 3D pose estimation problem by exclusively relying on the availability of a noisy kinematic signal and a 3D model of the instruments. While we can conclude that this was successfully demonstrated, several work directions should be considered in view of a potential real-world application, aimed at improving robustness and flexibility:

- **improved problem modelling**: the *physical* module ρ, consisting of the kinematic model r, the robot-camera transformation matrix b and the renderer model c, is here implicitly learnt from a synthetic dataset as the direct mapping between kinematics and tool projections. This strategy, while effective for the presented study, lacks in flexibility and, potentially, robustness. First of all, it is based on the assumption that the robot-camera transformation b is perfectly known, a hypothesis which may not be guaranteed in practice. More recent solutions like [DWLU23] include b as optimization target, allowing for its imprecise initialization. Alternatively, recent works have shown the feasibility of performing hand-eye calibration without calibration objects, more flexible to changes in robotic set-ups [PVES21]. A

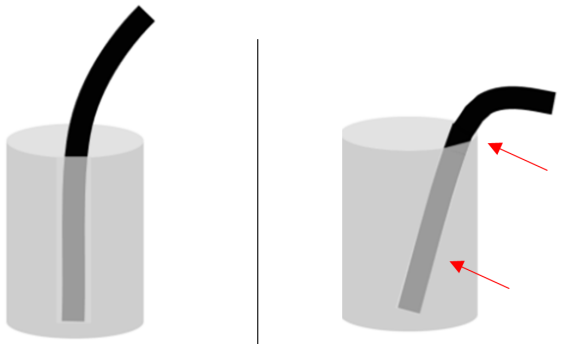

Figure 5.9: Left: modelled tool configuration according to a certain kinematic configuration $\mathbf{k}$*. Right: actual tool configuration due to unmodelled phenomena, like tool slackening and tool-channel interaction.*

similar assumption is also made for the camera parameters. While having a known calibration matrix is a common assumption for vision-based 3D pose estimation methods, our solution does not currently account for possible changes in such matrix, as this would require to completely retrain ρ. This severe limitation should be obviated by using differentiable renderering operations [KUH18], as done in recent works for both rigid and flexible surgical tools [DZK+22, LLGY23].
Finally, more specifically to the current STRAS robot setup, the current kinematic model is limited by a constant curvature assumption and by unmodelled phenomena like tool slacking inside the working channel and interaction of the tool with the channel border. This results in possible configurations which the current model is unable to match (Figure 5.9). Increasing model complexity, as shown for example in [Cab16], may bring direct benefits to our solution;

- **improved noise modelling**: in order to account for $\mathbf{k_s}$ uncertainty in the *backgroundizer*, we adopted the naive strategy of expanding the area covered by the tools according to their position in a square grid. This approach misses out on interesting opportunities to integrate additional prior knowledge. For example, the experimental observation of tool projection in the image space according to $\mathbf{k_s}$ suggests that certain tool configurations are more error-prone than others. This knowledge could be formalized, and injected into the problem to better model $\mathbf{k_s}$ uncertainty;

- **refining instead of regressing**: KI-BOT was designed to carry out the kinematic regression task directly from images. This choice has the advantage of not requiring real-time access to the robot API to read the kinematic data, uncommon in the operating room [AOH+18]. However, if recorded kinematics $\mathbf{k_s}$ is accessible in real-time, the regression problem could be reformulated as predicting the residual error vector $\hat{\varepsilon}_{\mathbf{k}}$ to compensate for the inaccuracy of $\mathbf{k_s}$. This would help ground model predictions to the measured kinematic values $\mathbf{k_s}$, minimizing the risk of inaccurate predictions, due for example, to tool occlusion;

- **improved regressor and decoder problem formulation**: the regressor model currently predicts a fixed number of joint values, describing the configuration of each tool potentially attached to the robotic system. If a tool is not visible in an

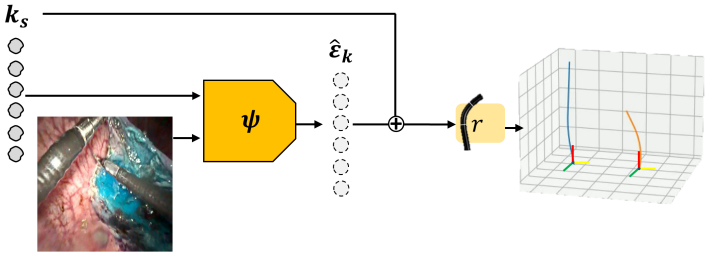

Figure 5.10: Potential inference architecture modification, if robot kinematics $\mathbf{k_s}$ *is available in real-time.*

image, the regressor learns to predict a kinematic configuration corresponding to an empty tool reprojection mask. However, a more efficient way to tackle this problem could be a multi-stage pipeline, where tools are first localized at the image level and then independently processed by the regressor, which would estimate the configuration of each visible tool independently. Furthermore, the decoder is not currently informed about the appearance of the tools, but only about their kinematic configuration. This is effective for the current set-up, where the two instruments have a similar and simple appearance, which can be implicitly learnt by the decoder as a prior. However, in order to deal with more complex instrument designs with significant inter-class tool variability, such as for the da Vinci® robotic system, additional information should be provided to the decoder.
These two limitations could be tackled by adding an instance segmentation module, predicting individual tool masks and classes, fed respectively to regressor and decoder models, in order to achieve instance-wise pose regression and class-informed decoding. A possible integration of KI-BOT with our PAF-IS approach for instance segmentation is shown in Figure 5.11. However, as discussed below, the availability of unsupervised methods for image-level localisation may facilitate KI-BOT training by replacing the image reconstruction objective;

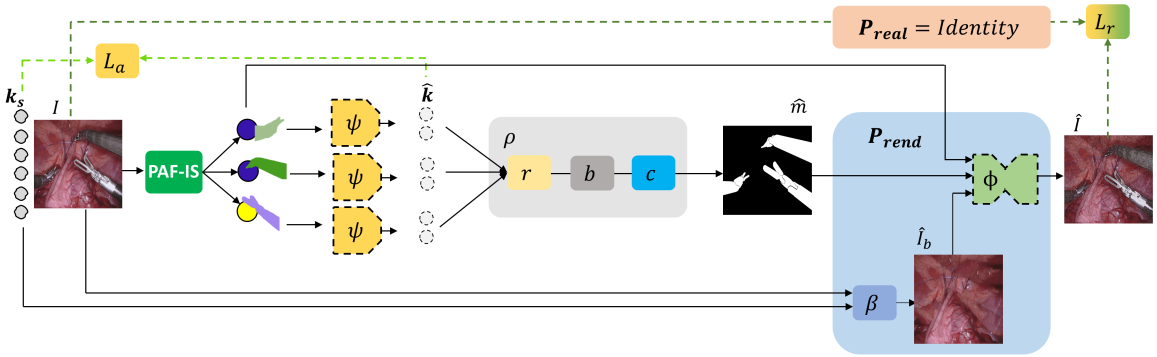

Figure 5.11: Potential training architecture modification integrating instance segmentation by PAF-IS to perform instance-wise 3D pose estimation and class-informed decoding.

- **integration of image-level localisation**: the image reconstruction-based objective was originally developed to avoid relying on manual annotations to learn the mapping operation P_{real}. While yielding satisfying results, such a strategy could now be potentially replaced by tools like FUN-SIS (Chapter 3), directly mapping the image

I into the same domain as $\hat{m}$, without requiring manual annotations. A potential KI-BOT architecture integrating FUN-SIS is shown in Figure 5.12, top. Interestingly, such architecture could be in principle applied in the laparoscopic domain, as it lifts the need for recorded kinematics (Figure 5.12, bottom). However, in the laparoscopic domain the transformation matrix b would be completely unknown a priori, because dependent on trocar placement. This could open interesting research directions, extending the task of 3D pose estimation beyond robotic surgery.

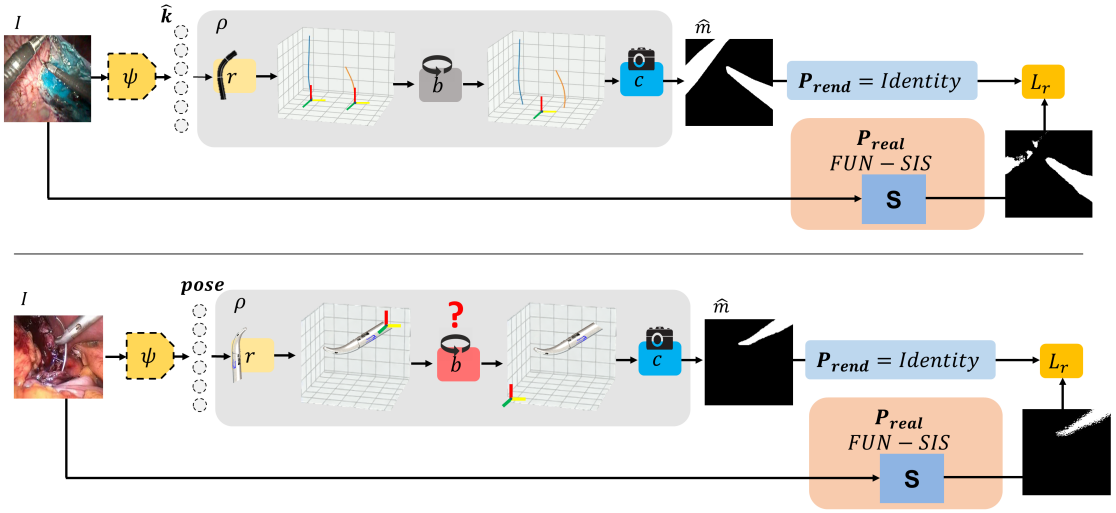

Figure 5.12: Top: potential training architecture modification integrating unsupervised tool segmentation by FUN-SIS as P_{real} mapping. The loss L_r could become a cross-entropy between projected and predicted masks. Bottom: application of such architecture to laparoscopic images. Such a solution would need to address the problem of the unknown robot-camera transformation b.

5.6 Conclusion

After tackling the tool localisation problem in the image space in Chapters 3 & 4, in this Chapter we explored the use of strong prior knowledge to extend localisation to the 3D space. The proposed KI-BOT approach aimed at showing the feasibility of tackling the 3D pose estimation problem by exclusively relying on the availability of a noisy kinematic signal and a 3D kinematic model of the instruments.
Overall this Chapter contributes to demonstrating how general problem knowledge, both cheaply available weak knowledge (e.g. kinematics, shape-priors), and strong repurposable knowledge (e.g. 3D kinematic modelling), can effectively replace manual annotations to tackle increasingly complex tool localisation tasks, from binary segmentation to 3D pose estimation.

CHAPTER 6

Conclusion

Contents

6.1 A Unified Framework for Learning from Unlabelled Datasets 134

6.1.1 Framework Contextualization 137

6.2 Discussion and Future Work 138

6.2.1 Limitations 138

6.2.2 Opportunities 139

6.2.3 Open Questions 140

This final Chapter recapitulates the contributions presented in this thesis and discusses them in view of further research work and translation. Section 6.1 presents a general framework collecting such contributions, highlighting their analogies and contextualizing them with respect to currently popular topics in the deep learning community. Current limitations, paths for further development and open questions are then discussed in Section 6.2.

6.1 A Unified Framework for Learning from Unlabelled Datasets

The goal of the work presented in this thesis was the development of methods for surgical instrument localisation and identification, not requiring manually annotated data to train. This was obtained by selecting suitable sources of information and designing frameworks able to digest such information, extracting effective supervision signals for deep learning model training. Regardless of the specific task tackled and prior/complementary knowledge source used, all the proposed contributions can be seen as instantiations of the general framework presented in Figure 6.1 (right). This framework ties together our contributions, and helps contextualize them with respect to relevant deep learning paradigms like unsupervised learning, self-supervised representation learning and learning-from-noisy-labels. Figure 6.1 recalls Figure 1.17 of Chapter 1 and describes how our work addressed our main research questions, posed in Section 1.4.4. The figure shows the standard fully-supervised learning framework (left) and the general framework adopted by our contributions to effectively learn from prior and complementary knowledge sources (right).

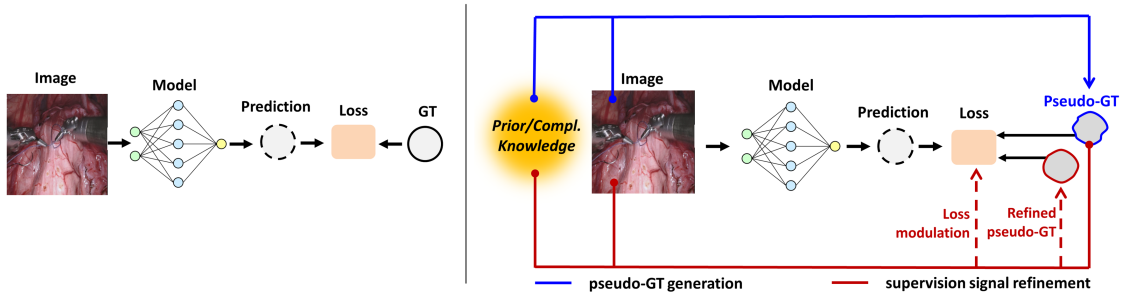

Figure 6.1: Left: fully-supervised learning framework. Right: general framework adopted in this work for model training with no manually annotated GT. The training follows two paths. Blue: pseudo-GT generation from unlabelled data and prior/complementary problem knowledge. Red: supervision signal denoising, actuated as direct pseudo-GT refinement or loss modulation. Such signal can be obtained from the pseudo-GT signal itself, the unlabelled data and additional prior/complementary knowledge.

In the absence of a manually annotated GT, each contribution, more or less explicitly fabricates a **pseudo-GT signal** from unlabelled data and prior/complementary knowledge (Figure 6.1,right, blue line). While inexpensive to obtain, such pseudo-GT is often noisy and would fail to effectively supervise the training of an output model. For this reason. a **denoising signal** is produced from the pseudo-GT itself, the unlabelled data and additional prior/complementary knowledge (Figure 6.1,right, red line): such signal

aims at improving the quality of supervision, either by directly refining the pseudo-GT or by modulating the loss between model's predictions and pseudo-GT. Each contribution can be seen as the application of this general framework to solve a specific task using a certain source of prior/complementary knowledge.

FUN-SIS, for binary segmentation, generates pseudo-GT masks from unsupervised segmentation of optical flow images, by incorporating tool shape-priors and relying on the different motion properties of tools and soft tissues (Figure 6.2, blue line).
It then modulates the loss between Student's network predictions and optical flow masks (pseudo-GT) by means of a localised binary Intersection-over-Union masking. This strategy exploits suitable noise properties of the pseudo-GT signal to automatically prevent back-propagation from potentially mislabelled pixels (Figure 6.2, red line);

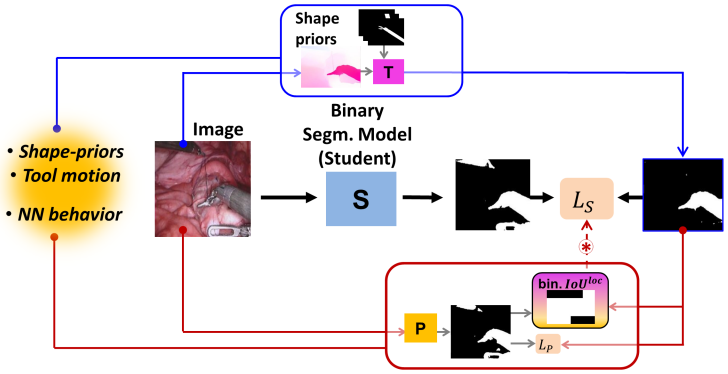

Figure 6.2: Overview of the FUN-SIS approach (Chapter 3) as the application of the general learning framework shown in Figure 6.1 (right).

PAF-IS, for tool instantiation, generates a pseudo-GT displacement field from Connected components instantiation of binary segmentation masks, relying on prior knowledge about tool positioning in the field-of-view (Figure 6.3, blue line).
Prior knowledge of tool positioning and laparoscopic triangulation is then exploited to automatically select potentially overlapping tools from the Connected Components instances (pseudo-GT) and discard them from training. This step cleans the supervision signal, but can potentially reduce the complexity of the training problem. Problem representativeness is recovered via artificial augmentation of the supervision signal, by pasting randomly selected instances from the training set (Figure 6.3, red line);

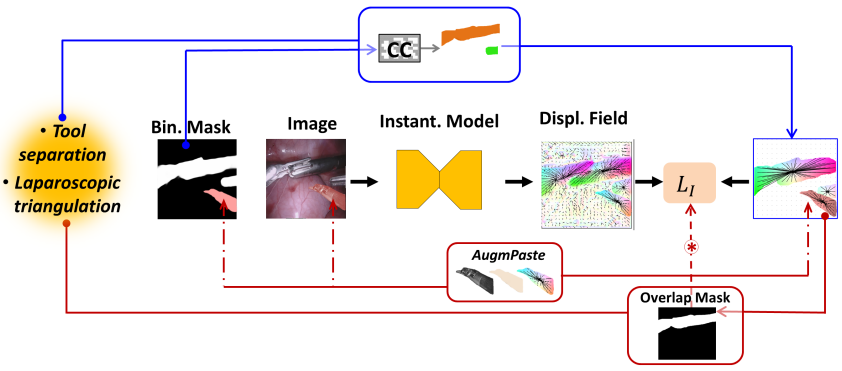

Figure 6.3: Overview of the PAF-IS approach for tool instantiation (Chapter 4) as the application of the general learning framework shown in Figure 6.1 (right).

PAF-IS, for instance classification, learns powerful feature representations via self-supervised learning, exploiting intrinsic temporal information of the video data. Training set features are then clustered, and each cluster is assigned to a prototype label, yielding a pseudo-GT signal to train the Student model for instance-wise classification (Figure 6.4, blue line).
The Student model is then trained exploiting prototype labels (pseudo-GT) as source of information to match instances with binary tool presence labels (Figure 6.4, red line);

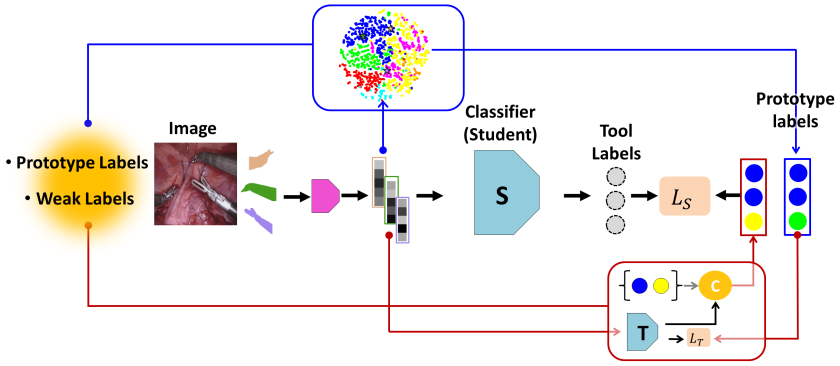

Figure 6.4: Overview of the PAF-IS approach for instance classification (Chapter 4) as the application of the general learning framework shown in Figure 6.1 (right).

KI-BOT, for 3D pose estimation, directly uses raw recorded kinematics to guide network optimization towards a space of plausible solutions (Figure 6.5, blue line).
Since the recorded kinematics is usually too inaccurate to directly supervised the kinematic regressor model training, an additional loss is computed. To this aim, the image itself is used as target to compute the image-based reconstruction loss L_r for regressor network optimization. This is achieved by incorporating the 3D kinematic model of the instruments in the architecture, allowing a differentiable mapping between kinematics and image-space tool projection. The projected tool image is combined with the *backgroundized* frame, obtained from the raw kinematics (pseudo-GT), and used to reconstruct the input frame for loss computation (Figure 6.5, red line).

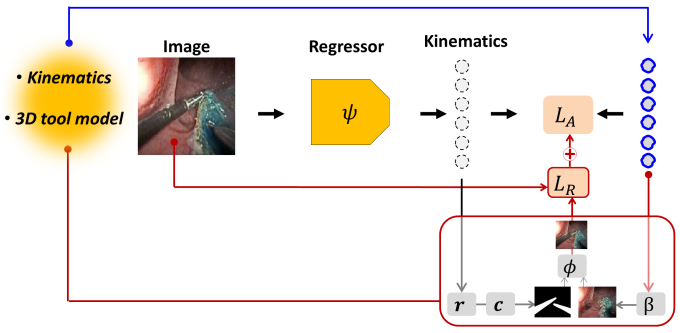

Figure 6.5: Overview of the KI-BOT approach for 3D pose estimation (Chapter 5) as the application of the general learning framework shown in Figure 6.1 (right).

For all the presented contributions, the denoising signal originates from the pseudo-GT, but incorporates additional information to solve the problem (e.g. additional labels for PAF-IS instance classification, 3D tool modelling for KI-BOT). This leads to a signifi-

cant improvement in the quality of the supervision signal, whose impact on each of our contributions can be seen in Table 6.1.

	FUN-SIS	PAF-IS		KI-BOT
Task	BS [IoU]	TI [mAP]	IC [IoU]	3D [IoU]
pseudo-GT	40.08	56.86	42.33	43.40
final model	83.77 (+109.0%)	63.70 (+12.0%)	52.44 (+23.9%)	55.40 (+27.6%)

Table 6.1: Comparison between pseudo-GT quality and final model's performance, after training on the denoised signal. Results for our contributions on the tasks of binary segmentation (BS), tool instantiation (TI), instance classification (IC) and 3D pose estimation (3D). The reported metrics were taken from the main results tables in Chapters 3, 4 & 5.

6.1.1 Framework Contextualization

The general framework shown in Figure 6.1 allows to organically link the work proposed in this thesis to several important deep learning paradigms and to highlight its transversal contribution to the community:

- **learning-from-noisy-labels**: as the generated pseudo-GT is often noisy, the proposed solutions to refine it can be seen as ways to effectively learn from noisy labels. Such a problem has been largely explored in the general deep learning community, motivated by the growing availability of platforms to cheaply collect large quantities of annotations at the price of potential inaccuracies (e.g. crowd-sourcing platforms, reliable web search engines). However, research work in the surgical computer vision community is still limited. This can be due to the reduced availability of noisy labels in the field, where the use of tools like crowd-sourcing platforms to obtain labels is less common. The work presented in this thesis shows the potential value of exploring the problem of learning-from-noisy-labels for tool localisation and identification tasks, where noisy labels can be cheaply obtained directly from data and prior knowledge.

- **self-supervised representation learning**: as shown by recent works in the surgical computer vision domain, self-supervised pre-training, in combination with a reduced set of labelled data, can boost model performance and reduce training time. In this thesis, we integrated self-supervised representation learning as part of PAF-IS feature learning step, for instance segmentation. Self-supervised learning was used to learn powerful instance-wise representations, which we clustered together for prototype labelling. This step dramatically minimized human annotation effort to less than the 0.2% of the fully-supervised case, significantly less than what is commonly required by semi-supervised approaches. This application further reinforces the potential value of self-supervised learning: beyond pre-training for semi-supervised approaches, representation learning can play an important role to focus annotation effort to a few prototype samples, maximizing annotation efficiency.

- **unsupervised learning**: the term *unsupervised learning* has been used in the deep learning community with different meanings. Early works mainly used it to define clustering and dimensionality reduction techniques [GCB04]; more recently the term has become an alias for self-supervised representation learning [RS21].

Furthermore, in the surgical computer vision community, different works have borrowed the term to refer to approaches aimed at learning specific tasks, like instrument segmentation, without requiring manual annotations [PSN20, LWJ+20a]. Similarly, in this thesis we adopted the term to define all those approaches whose training is not bottlenecked by the need for manually annotated labels paired with training samples. This condition is essential to break the linear relationship between datasets size and cost of annotations, allowing to potentially increase the first one without affecting the second one. We therefore believe that these approaches should fall under the definition of *unsupervised learning*. Nonetheless, we think that an improved and more granular definition of learning paradigms could help the community align on such a relevant topic, as we further discuss in Section 6.2.3.

6.2 Discussion and Future Work

This final Section first discusses potential lines of work originating from this thesis, either aimed at addressing current limitations (Section 6.2.1) or at exploiting the opportunities provided by our contributions (Section 6.2.2). Open questions arising from this work are then presented in Section 6.2.3.

6.2.1 Limitations

From a methodological stand-point, individual limitations and potential future work directions, specific to each contribution, have been discussed at the end of Chapters 3, 4 & 5. We now discuss two more general limitations, related to the use of prior and complementary knowledge sources, that future work could address:

- **limited effective usage of the available information**: the proposed frameworks were designed to convert unorthodox sources of information into suitable supervision signals for deep learning model training. However, during this conversion, part of the information may be lost. For FUN-SIS for example, shape-priors were only used as part of the pseudo-GT generation. Shape consistency information could have been injected at the following stages, for example, to regularize Proxy and Student predictions. In KI-BOT, raw kinematics was used to *backgroundize* frames accounting for a systematic error, unaware of temporal information or error-prone configuration. PAF-IS instantiates tools only based on tool positioning on the field-of-view, overlooking motion information which could potentially disambiguate the cases of overlap. Understanding how to maximize the effective usage of the available information is a necessary next step of the proposed work that future research could investigate.

- **limited representativeness of validation**: in order to compare our solutions with state-of-the-art methods, training and validation were carried out on popular benchmark datasets like EndoVis. Here, fully-supervised approaches still hold a certain performance gap with respect to our solutions. However, as discussed in Section 1.4.2, the representativeness of such datasets is often questionable, as their size is limited by the cost of annotations. The development of large multi-centric

and multi-procedural datasets for training and validation is therefore an imminent requirement for the surgical data science community. This can enable a more representative benchmarking of the developed approaches, clearly showing the right steps to advance toward clinical translation. Moreover, the metrics commonly adopted for evaluation are completely agnostic with respect to the downstream clinical applications the developed solutions should serve. The need for metrics specifically designed for clinically-oriented validation is further discussed as part of open questions in Section 6.2.3.

6.2.2 Opportunities

The value of the proposed contributions is a direct consequence of their ability to train without manual annotations, enabling instrument localisation and identification on data coming from completely unlabelled domains. We now present application opportunities for our solutions, pointing out how they can mitigate the main negative consequences of the *annotation bottleneck*, described in Section 1.4.2:

- *contribute to the development of clinically valuable applications/studies*: as described in Section 1.3.2, a wide range of applications for computer-assisted surgery and surgical data science depends on tool localisation and identification. For such applications, the need for tool related information can represent a bottleneck preventing extensive validation and deployment. For example, in order to evaluate their approach for automatic surgical skill assessment, [LZK+21] were required to perform extensive data labelling for skill grading (the final goal) and tool detection (enabling the study). The effort dedicated to tool annotation, could have been highly mitigated by the availability of tools like PAF-IS, and redirected towards a more extensive skill grading (more surgeons, more procedures etc.), enlarging the scope of such research and its clinical value. Similarly, for different works on 3D scene reconstruction [WLFD22], augmented reality [TPPV21] and visual-servoing [ZRCM+21], our solutions could help enlarge the size of methodological validation, accelerating their deployment;

- *facilitate benchmark datasets creation*: as already discussed, building large and representative annotated datasets for models benchmarking is a crucial need for the whole surgical data science community. AI-powered annotation platforms are gaining interest for medical image analysis [PWA+19]. Such platforms can provide predictions from pre-trained AI models, to be refined by the annotator, significantly speeding up the annotation process. Tools like FUN-SIS and PAF-IS could be integrated into such annotation platforms, facilitating the annotation of large sets of unlabelled data for the instrument segmentation task;

- *promote task de-compartmentalization*: the availability of tools like FUN-SIS, PAF-IS and KI-BOT can encourage researchers in the surgical computer vision community to tackle relevant tasks like phase/step segmentation and action-triplet recognition, relying on instrument-related information. As discussed in Section 1.4.2, this is normally prevented by the lack of datasets extensively annotated for multiple tasks. As an example, action-triplet recognition approaches could integrate tool segmentation information provided by PAF-IS, to guide target

identification. Beyond simple information integration, the general framework presented in Figure 6.1 (right) could be applied to different tasks. For example phase segmentation could be tackled by generating a pseudo-GT signal from binary tool presence information, refined by prior knowledge on phase casual relationships. Overall, this could open new interesting research directions, promoting the de-compartmentalization of surgical computer vision research.

6.2.3 Open Questions

The work presented in this thesis tried to provide answers to the research question asked at the end of Chapter 1: *can a deep learning model effectively learn from general knowledge about surgical tools?*
While we can conclude that our work positively answered this question, other important and urgent questions arise from it. Such questions are listed below and briefly commented on, leaving the reader free to form his own opinion.

Is the standard protocol *collect data-annotate-train* obsolete?
As shown by the work presented in this thesis, the fast progress of self-supervised learning is providing powerful tools to learn from unlabelled data. Yet, the common approach to deep learning applications development starts with the systematic annotation of the available data, usually relying only on qualitative intuitions from the annotators to decide *which* of the available data should be annotated and *how*. As the interest of the community in a more Data-Centric AI is increasing [ZBL+23], such an approach may need to be reevaluated. As shown in this thesis the information gathered through self-supervised learning can be directly used to target annotation efforts on valuable samples, letting the data themselves reveal the optimal annotation strategy.

Is it time to redefine learning paradigms?
Despite an increasingly growing interest in alternatives to full-supervision, learning paradigms ontology still relies on a few categories, like *unsupervised*, *semi-supervised*, *weakly-supervised* learning. These terms are usually too broad to convey information about the amount of supervision required for model training, and the annotation effort needed to obtain it. As the language shapes the way we think [Bor18], the lack of adequate terms can have the long-term effect to reduce or distort the relevance of such topics. A more transparent way to define learning paradigms could start from an objective quantification of the amount of supervision each method requires, based on specific criteria: *does the method require manual annotations? If yes, do annotations need to be paired with input samples? Does every input sample need to be labelled? What type of manual annotations are required (video-level, frame-level, pixel-level)? If labels can be automatically obtained, do they need ad-hoc setups like simulators? Are these setups readily available to the community and at what cost? Is their availability confined to specific types of surgery (e.g. robot-assisted)?* Answering these questions may allow a finer annotation cost description, more oriented towards clinical translation.

What is the performance level we should strive for?
Ideally, the overarching goal for surgical computer vision research should be the development of models enabling the deployment of clinically valuable applications.

Available metrics to evaluate such models are commonly borrowed from the general computer vision community, and are therefore agnostic with respect to surgical applications. The segmentation task, for example, is commonly evaluated using Intersection-over-Union score (IoU). However, different down-stream applications can have significantly variable requirements over the quality of segmentation masks. Augmented reality applications using tool masking may require a perfect boundary segmentation of the whole instruments. A surgical skill-assessment approach may have softer requirements on boundary segmentation quality, as long as instruments are correctly localised. An action-triplet recognition model integrating instrument segmentation may just need a good localisation of instruments' tip, as this is usually the part carrying out the action. IoU metric does not reflect these different needs. If the available tools to measure performance are not reflective of the final application needs, how can we know when computer vision models will be ready for translation?

How do the proposed unsupervised solutions conform to regulatory guidelines?
Regulatory guidelines are quickly evolving to keep up with the latest technological advancements. Artificial Intelligence, and in particular deep learning, challenge standard regulatory guidelines due to their black-box nature, and their adaptive behaviour, i.e. the possibility to retrain models as new data are collected. Recently, The U.S. Food and Drug Administration (FDA), Health Canada, and the United Kingdom's Medicines and Healthcare products Regulatory Agency (MHRA) have jointly identified a list of guiding principles that can inform the development of Good Machine Learning Practice (GMLP) [FA+21]. In such a guideline, a great emphasis is placed on the representativeness of both training and testing data. Point 3 of the list reads *"Data collection protocols should ensure that the relevant characteristics of the intended patient population (for example, in terms of age, gender, sex, race, and ethnicity), use, and measurement inputs are sufficiently represented in a sample of adequate size in the clinical study and training and test datasets, so that results can be reasonably generalized to the population of interest"*. Unsupervised solutions not requiring manual annotations to train, such as the ones proposed in this thesis, can certainly help deep learning approaches conform to such requirements, by extending the size of training sets to reduce selection bias. On the other side, solutions like FUN-SIS use mechanisms like loss modulation, aimed at improving the quality of the supervision signal by reducing the effective size of the training sets. These approaches may require specifically designed protocols to ensure that the *effective* training data are free from biases.

Appendices

APPENDIX A

List of Publications

Journal Articles

- **Sestini, L.**, Rosa, B., De Momi, E., Ferrigno, G., & Padoy, N.

  A kinematic bottleneck approach for pose regression of flexible surgical instruments directly from images

  IEEE Robotics and Automation Letters, 6(2), 2938-2945, 2021

- **Sestini, L.**, Rosa, B., De Momi, E., Ferrigno, G., & Padoy, N.

  FUN-SIS: A fully unsupervised approach for surgical instrument segmentation

  Elsevier Medical Image Analysis, 102751, 2023

- Ramesh, S., Srivastav, V., Alapatt, D., Yu, T., Murali, A., **Sestini, L.**, Nwoye I., C., Hamoud, I., Sharma, S., Fleurentin, A., Exarchakis, G., Karargyris, A. & Padoy, N.

  Dissecting Self-Supervised Learning Methods for Surgical Computer Vision.

  Accepted for publication in Elsevier Medical Image Analysis, 2022

- Mascagni, P., Alapatt, D., **Sestini, L.**, Altieri, M. S., Madani, A., Watanabe, Y., Alseidi, A., Redan A., J., Alfieri, S., Costamagna, G., Boškoski, I., Padoy N. & Hashimoto, D. A.

  Computer vision in surgery: from potential to clinical value

  npj Digital Medicine, 5(1), 163, 2022

- **Sestini, L.**, Rosa, B., De Momi, E., Ferrigno, G., & Padoy, N.

  PAF-IS: a Pixel-wise Annotations Free framework for Instance Segmentation of surgical tools

  In preparation for submission

Conference Presentations

- S.Herrera, J. F. G., Pore, A., **Sestini, L.**, Sahu, S. K., Liao, G., Zanne, P., Dall'Alba, D., Hernansanz, A., Rosa, B., Nageotte, F. & Gora, M.

  Autonomous image guided control of endoscopic orientation for OCT scanning

  Proceedings of Conference on New Technologies for Computer and Robot Assisted Surgery (CRAS), 2022

Long Abstracts

- **Sestini, L.**, Rosa, B., De Momi, E., Ferrigno, G., & Padoy, N.

  Unsupervised Binary Instrument Segmentation Using Prior Instrument Knowledge

  To be presented at the International Conference on Information Processing in Computer-Assisted Interventions (IPCAI), 2023

APPENDIX B

Résumé en français

B.1 Introduction

Alors que nous entrons dans une ère d'évolution technologique rapide alimentée par le big data et la science des données, la chirurgie est vouée à être radicalement transformée. La salle d'opération d'aujourd'hui, le centre des soins chirurgicaux, est un environnement technologique et numérique, où la variété des signaux produits par les appareils numériques peut capturer et "datafier" la complexité des soins intra-opératoires aux patients (Figure B.1).

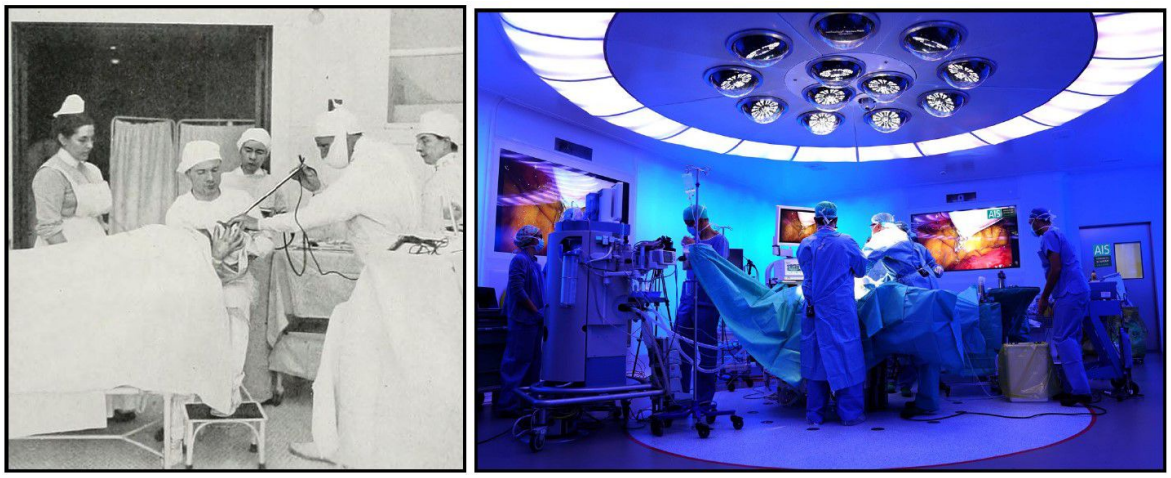

Figure B.1: Évolution de la salle d'opération.

L'analyse de ces données par le biais de l'intelligence artificielle, qui ne cesse de se développer, offre des possibilités apparemment infinies pour rendre la chirurgie plus sûre et la démocratiser. L'aide à la décision en temps réel, la réalité augmentée, l'évaluation automatique des compétences chirurgicales, la documentation postopératoire objective et ciblée ne sont que quelques exemples de solutions basées sur les données qui pourraient transformer la chirurgie d'un métier artisanal, basé sur les connaissances indi-

viduelles des chirurgiens, en une discipline décisionnelle objective, basée sur l'échange et l'interprétation continus de données. Toutefois, le traitement et la monétisation appropriés de volumes de données aussi importants sont loin d'être triviaux. Les vidéos endoscopiques, la source de données la plus riche et la plus complète pour étudier l'acte chirurgical en chirurgie mini-invasive, sont généralement trop volumineuses pour être stockées, et trop peu structurées pour être directement utilisées pour des applications en aval. La vision par ordinateur, alimentée par l'apprentissage profond, offre des outils puissants pour extraire des informations pertinentes à partir de données vidéo brutes. En particulier, la localisation et l'identification automatiques des instruments chirurgicaux à partir de ces vidéos est une composante essentielle d'applications en aval précieuses telles que l'évaluation automatique des compétences chirurgicales et l'aide à la décision en temps réel, visant à faciliter la formation chirurgicale et à fournir une assistance intra-opératoire. La plupart des solutions disponibles pour résoudre ce problème utilisent des approches d'apprentissage entièrement supervisées pour former des modèles d'apprentissage profond sur des données annotées manuellement. En raison du coût des annotations, l'entraînement de ces modèles est limité à des ensembles restreints de données étiquetées et curatées, ce qui peut avoir un impact sur leur capacité de généralisation sur des données réelles.

B.2 Motivation

B.2.1 Le goulot d'étranglement de l'annotation

Comme indiqué dans la section précédente, l'apprentissage profond (Deep Learning - DL) est la méthode actuellement privilégiée pour résoudre le problème de la localisation et de l'identification des instruments. Plus généralement, au cours de la dernière décennie, l'apprentissage profond, alimenté par une quantité sans cesse croissante de données et de capacités informatiques, a surpassé les algorithmes de vision par ordinateur (Computer Vision - CV) standard dans une variété de tâches telles que la détection d'objets, la segmentation d'images et la classification. Par rapport aux algorithmes CV standard, qui nécessitent une sélection et une extraction manuelles des caractéristiques, le DL permet à une fonction de réseau neuronal d'apprendre ses propres paramètres à partir d'un ensemble de données d'apprentissage, sans être explicitement programmée. Dans le paradigme largement utilisé de l'apprentissage entièrement supervisé, ce résultat est obtenu en optimisant les paramètres du réseau neuronal avec l'objectif avide de produire des prédictions correspondant à une vérité de terrain annotée manuellement. Bien que ce paradigme ne soit pas exempt de défauts, tels que la possibilité d'un surajustement des données d'apprentissage, il a rapidement conduit à des résultats révolutionnaires dans le domaine de la CV. Dans le domaine de la CV chirurgicale, le processus mental suivi par les experts pour exécuter les tâches est le fruit d'années d'expérience et il est souvent difficile de le formaliser explicitement: la définition explicite des caractéristiques pertinentes pour le développement d'algorithmes de CV standard est donc prohibitive. Motivée par ce défi et alimentée par l'enthousiasme des premiers résultats, l'approche entièrement supervisée a été largement utilisée pour la plupart des tâches de CV chirurgicales, y compris la localisation et l'identification des instruments chirurgicaux. Cependant, cette approche unique, qui nécessite des annotations manuelles pour résoudre chaque problème de CV chirurgical, semble aujourd'hui insoutenable et incompatible avec la traduction

clinique. Pour garantir la robustesse et la capacité de généralisation, les approches DL doivent en effet être entraînées sur de grandes quantités de données, en tenant compte de la variabilité potentielle des données du monde réel. Cependant, à mesure que la taille des ensembles de données augmente, le coût de l'annotation s'accroît de façon linéaire. S'il n'est pas correctement traité, ce goulot d'étranglement de l'annotation peut sérieusement limiter les avantages que le big data chirurgical pourrait apporter à la chirurgie, en ayant un impact négatif sur la recherche et la traduction de différentes manières:

- manque de capacité de généralisation des modèles: l'effet le plus immédiat de la formation sur de petits ensembles de données est l'incapacité des modèles à donner de bons résultats sur des données inédites, en raison de facteurs de biais indésirables présents dans les données collectées. Cela peut poser de graves problèmes pour la traduction des technologies développées;

- représentativité limitée des ensembles de données de référence: la nature hautement compétitive de la recherche pousse les chercheurs à développer des solutions visant à surpasser les approches de pointe sur des ensembles de données de référence spécifiques. Ces ensembles de données permettent une évaluation normalisée des méthodes, favorisant ainsi une comparaison équitable entre elles. Toutefois, lorsque le nombre et la taille de ces ensembles de données de référence sont réduits en raison du goulet d'étranglement de l'annotation, leur capacité à représenter des problèmes réels peut être entravée. Cela peut conduire à l'élaboration de méthodes sur-optimisées pour obtenir de bons résultats sur des

  ensembles de données potentiellement non représentatifs, ce qui ne permet pas de faire progresser efficacement l'état de l'art;

- le cloisonnement des tâches: il est rare que l'on réannote des ensembles de données existants pour différentes tâches de CV chirurgicale. Cela peut avoir pour effet à long terme de cloisonner les tâches de CV chirurgical, car les chercheurs sont contraints de trouver des solutions aux problèmes sans s'appuyer sur des informations provenant de différentes tâches. Cela est extrêmement contre-intuitif pour les tâches de CV chirurgical, qui abordent souvent le même problème (par exemple, l'analyse du flux de travail chirurgical) sous différents angles (par exemple, la reconnaissance des gestes, l'identification des triplés d'action, la segmentation des phases/étapes), tout en bénéficiant potentiellement de la disponibilité des mêmes informations (par exemple, les informations relatives aux outils).

B.2.2 Possibilités de connaissances antérieures et complémentaires

Contrairement aux annotations manuelles standard utilisées pour aborder les tâches de localisation et d'identification des outils (par exemple, les boîtes de délimitation, les masques de segmentation), les connaissances préalables et complémentaires peuvent fournir des informations moins coûteuses et plus souples sur le problème. Les connaissances préalables, en particulier, ne sont généralement pas directement liées à un ensemble spécifique de données. Les informations sur la forme des outils et la distribution des couleurs, par exemple, peuvent être appliquées à tous les domaines chirurgicaux, pour des données collectées à partir de différentes procédures, réalisées avec différentes techniques et dans différents centres. Les connaissances

complémentaires, bien qu'elles soient généralement liées à un ensemble spécifique de données, peuvent souvent être obtenues automatiquement (par exemple, les données cinématiques et l'utilisation des outils par les systèmes robotiques), ou avec moins d'efforts d'annotation (par exemple, la présence d'outils binaires et les étiquettes de phase/étape) par rapport aux annotations manuelles standard. Ces caractéristiques font des connaissances préalables et complémentaires une source d'information plus générale et réutilisable, qui peut permettre aux modèles d'apprendre à partir de données non étiquetées, sans se heurter au problème du goulot d'étranglement de l'annotation. En revanche, l'intégration de ces informations dans les architectures DL standard n'est pas simple, par rapport à l'apprentissage entièrement supervisé standard.

B.2.3 Objectif de la recherche

Dans cette thèse, nous étudions les problèmes de localisation et d'identification des instruments chirurgicaux, à la fois au niveau de l'image, sous forme de segmentation, et dans l'espace 3D, sous forme d'estimation de la pose 3D. Motivés par la nécessité de supprimer les annotations manuelles, les connaissances préalables et complémentaires sur les outils - moins coûteuses et plus faciles à obtenir que les annotations manuelles - sont incorporées dans les architectures développées. Cela permet d'obtenir des approches sans annotation, qui peuvent être entraînées sur des données non étiquetées. La Figure B.2 présente une vue d'ensemble des contributions proposées.

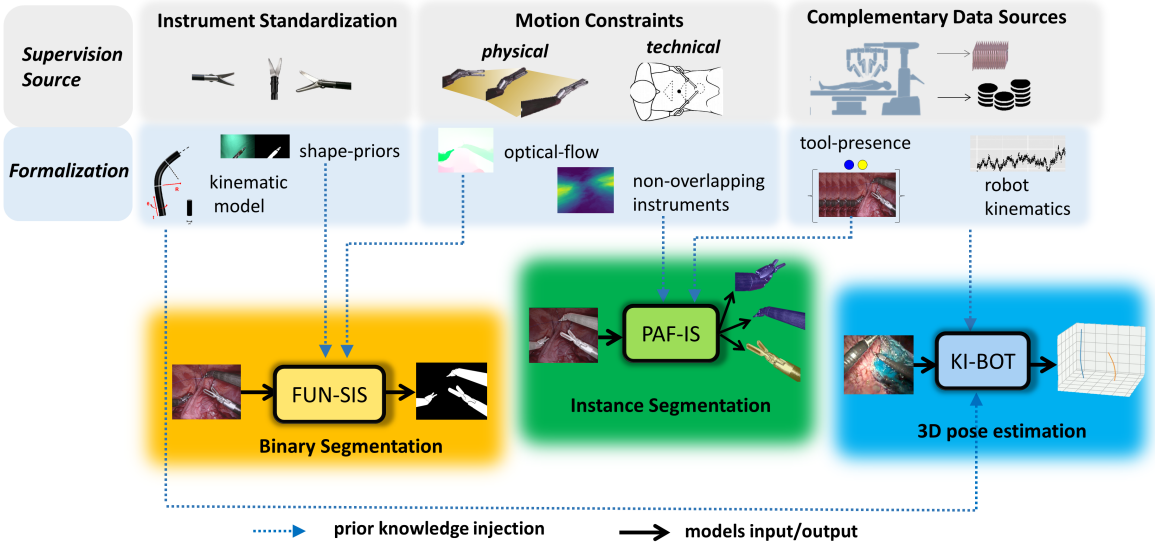

Figure B.2: Aperçu des principales contributions de la thèse.

B.3 FUN-SIS: une approche entièrement non supervisée pour la segmentation des instruments chirurgicaux

Notre première contribution est la conception d'une approche pour la segmentation binaire non supervisée des instruments. La solution proposée s'entraîne sur des vidéos totalement non étiquetées, en exploitant les connaissances préalables sur l'apparence

et le mouvement des instruments. Les connaissances préalables sur l'apparence des instruments sont formalisées sous la forme de masques de forme, de masques de segmentation binaire d'instruments chirurgicaux, que l'on peut obtenir de différentes manières, par exemple en recyclant des annotations existantes provenant de différents ensembles de données ou en projetant des modèles d'outils en 3D dans l'espace image. Pour le mouvement des instruments, une hypothèse simple est formulée: par rapport à l'anatomie environnante, les instruments chirurgicaux se déplacent de manière cohérente, c'est-à-dire que deux points proches d'un instrument se déplacent normalement dans la même direction. L'approche FUN-SIS (Figure B.3) est une méthode en trois étapes qui effectue une segmentation non supervisée des outils chirurgicaux sur des images de flux optique (étape I) et entraîne ensuite un modèle de segmentation par image sur les pseudo-étiquettes bruyantes générées à l'étape I à l'aide d'une nouvelle stratégie d'apprentissage à partir d'étiquettes bruyantes (étapes II et III).

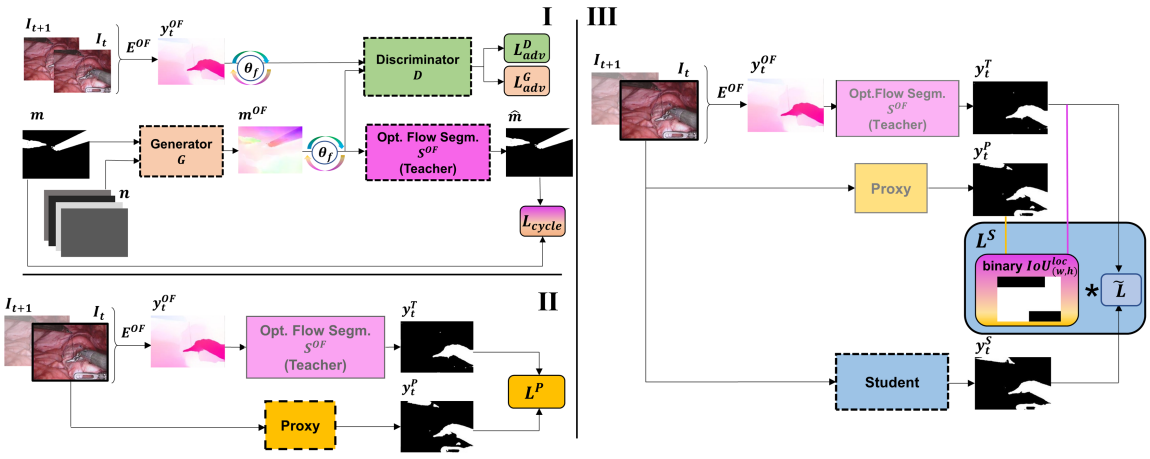

Figure B.3: Vue d'ensemble de l'architecture FUN-SIS.

Étape I: tout d'abord, un modèle (appelé Enseignant) est formé pour effectuer la segmentation des outils dans les images de flux optique. Par rapport aux images endoscopiques brutes, le domaine du flux optique permet de minimiser la variabilité de l'apparence de l'outil et de l'arrière-plan. Cela nous permet de configurer le problème d'entraînement comme une traduction de domaine non apparié, entre les images de flux optique et les masques de prieurs de forme (Figure 3-I), que nous résolvons à l'aide d'une architecture cycle-GAN modifiée. Une fois entraîné, le modèle de l'enseignant est utilisé pour générer des masques pseudo-étiquetés pour les vidéos endoscopiques non étiquetées. En raison de la sous- optimalité de l'estimateur de flux optique et du réseau Teacher, les pseudo-étiquettes obtenues ont tendance à être très bruitées. Nous montrons néanmoins que leurs propriétés particulières en matière de bruit permettent de les exploiter dans les étapes suivantes, ce qui améliore considérablement la précision de la segmentation.
Étape II: les pseudo-étiquettes générées à l'étape I sont utilisées pour superviser directement l'apprentissage d'un modèle (appelé Proxy) sur la segmentation des outils chirurgicaux sur des images individuelles. L'efficacité de cette étape est garantie par une propriété particulière du bruit affectant les pseudo-étiquettes, à savoir l'imprévisibilité. Le bruit affectant les pseudo-étiquettes, obtenu à partir d'une segmentation par flux optique pur, ne peut pas être appris par le réseau Proxy, qui traite une seule image à la fois et ignore donc le mouvement de l'outil, qui est en fin de compte responsable du

bruit. Comme le suggèrent plusieurs travaux sur l'apprentissage des réseaux neuronaux sous supervision bruyante, le réseau Proxy tente d'adapter les données à un modèle facile compatible avec les pseudo- étiquettes, c'est-à-dire en séparant les outils de l'anatomie, ce qui améliore considérablement les résultats de segmentation du réseau Enseignant.
Étape III: le réseau Proxy génère des masques de segmentation globalement corrects (du fait de l'apprentissage d'un modèle partagé facile), mais potentiellement inexacts au niveau local, car l'apprentissage ne peut pas converger vers une solution stable (ce qui affecte principalement les pixels difficiles à classer, tels que les pixels limites). Nous exploitons cet aspect pour améliorer les résultats de la segmentation en utilisant les prédictions du réseau Proxy pour sélectionner des régions probablement bien étiquetées des pseudo-étiquettes, sur lesquelles un troisième modèle, le réseau Student, est entraîné. Cette étape repose sur l'observation d'une deuxième propriété de bruit particulière des pseudo-étiquettes, la propriété de polarisation, qui décrit le fait que les outils chirurgicaux individuels sont généralement soit bien segmentés, soit complètement mal étiquetés par la segmentation du flux optique, en raison de leur mouvement constant. La sélection des régions est effectuée au moyen d'une version locale nouvellement introduite de la célèbre métrique Intersection-sur-Union (IoU), à savoir l'IoU local. L'intersection sur l'union locale mesure l'accord local entre les réseaux Proxy et Enseignant, ce qui permet de ne sélectionner que les régions où les deux réseaux sont d'accord. Par conséquent, le réseau de l'élève est formé sur un sous-ensemble des données disponibles à l'origine (les régions probablement bien étiquetées), pour lesquelles les pseudo-étiquettes sont généralement de bonne qualité. Cela permet de réduire considérablement le rapport de bruit effectif du signal de supervision sur lequel le réseau de Student est entraîné, ce qui permet une meilleure convergence et, en fin de compte, des résultats de segmentation de haute qualité.

L'approche proposée a été largement validée sur différents ensembles de données chirurgicales (procédures endoscopiques, robotiques et laparoscopiques flexibles). Sur l'ensemble de données populaire MICCAI 2017 EndoVis Robotic Instrument Segmentation Challenge, l'approche non supervisée proposée est presque aussi performante que les modèles entièrement supervisés de pointe. Les diagrammes en boîte présentés à la Figure B.4 montrent l'amélioration significative du réseau Proxy par rapport au réseau Teacher, et l'effet d'affinage du réseau Student, produisant des masques de segmentation précis et nets, ayant un IoU moyen avec les masques de vérité terrain seulement 5,22% en dessous de celui de la ligne de base entièrement supervisée. En plus des expériences principales, plusieurs études d'ablation ont été réalisées, mettant en évidence des aspects cruciaux de cette méthode tels que les exigences peu contraignantes en matière de quantité et de type de prieurs de forme, et l'efficacité de la stratégie proposée d'apprentissage à partir d'étiquettes bruyantes pour nettoyer le signal de supervision.

B.3.1 PAF-IS: un cadre sans annotation de pixels pour la segmentation d'instances d'outils chirurgicaux

Notre deuxième contribution consiste à proposer un nouveau cadre pour l'apprentissage de modèles de segmentation d'instances, conçu pour minimiser l'effort d'annotation humaine en supprimant le besoin d'annotations sémantiques et d'instances au niveau du pixel. Sans signal de supervision explicite, notre solution apprend à extraire des

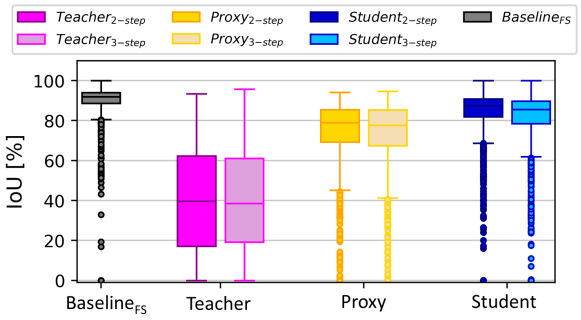

Figure B.4: Résultats de FUN-SIS sur l'ensemble de données EndoVis 2017, et comparaison avec la ligne de base entièrement supervisée.

instances d'outils individuels à partir de masques de segmentation binaires, et obtient, pour chaque instance d'outil, une représentation puissante des caractéristiques via l'apprentissage contrastif auto-supervisé. Ces représentations par instance guident la sélection automatique d'un petit nombre d'instances (aussi peu que 8 dans nos expériences), présentées à un utilisateur humain potentiel pour l'étiquetage du type d'outil. Les informations recueillies, en combinaison avec les étiquettes binaires de présence d'outil, guident l'apprentissage d'un classificateur par instance, prédisant une étiquette de type d'outil pour chaque instance d'outil. Les trois étapes de l'apprentissage illustrées à la Figure B.5 sont maintenant décrites.

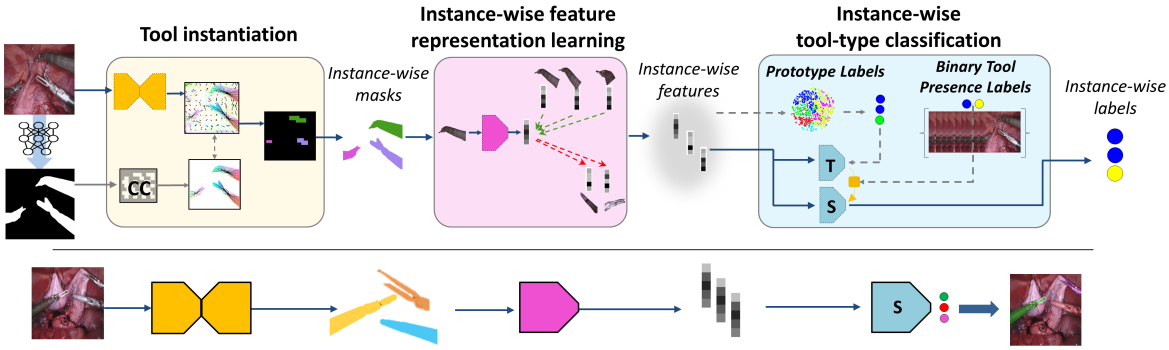

Figure B.5: Vue d'ensemble de l'architecture PAF-IS.

Instanciation des outils: en l'absence d'instanciation de la vérité de terrain, nous partons de l'hypothèse que les chirurgiens ont tendance à éviter le chevauchement des instruments chirurgicaux, afin de réduire les risques d'occlusions mutuelles et d'interactions indésirables entre les outils. Par conséquent, nous fabriquons un signal de pseudo-supervision à partir de l'instanciation de la composante connectée des masques de segmentation binaires. Ce signal de supervision est ensuite affiné en fonction du positionnement de l'instrument dans l'espace de l'image afin d'entraîner efficacement un modèle d'instanciation, produisant un masque de segmentation pour chaque outil individuel.

Apprentissage de la représentation des caractéristiques par instance: en l'absence d'étiquettes sémantiques par pixel, nous apprenons, pour chaque instance d'outil, une représentation des caractéristiques décrivant son contenu sémantique. Pour ce faire, nous nous appuyons sur les informations temporelles intrinsèques des séquences vidéo. Plus précisément, nous concevons une approche d'apprentissage contrastive basée sur

le suivi d'instances locales pour tirer des échantillons positifs et négatifs. Cette étape permet d'obtenir de puissantes représentations de caractéristiques au niveau de l'instance, fournissant les informations nécessaires pour résoudre l'étape de classification finale; Classification par instance: En l'absence d'étiquettes sémantiques au niveau du pixel, nous nous appuyons sur les étiquettes binaires de présence d'outil pour résoudre cette tâche. Cependant, pour exploiter ces informations, chaque instance d'outil doit être associée à une étiquette binaire de présence d'outil. À cette fin, nous injectons une quantité minimale de connaissances humaines, spécifiquement collectées pour maximiser leur contenu informatif et les exploiter pour résoudre la tâche de mise en correspondance. Plus précisément, nous sélectionnons automatiquement un petit nombre d'instances hautement représentatives (instances prototypes) et demandons à un utilisateur humain de les étiqueter. Les informations recueillies sont ensuite utilisées pour faire correspondre les étiquettes de présence d'outils binaires et les instances, en exploitant une nouvelle approche enseignant-étudiant, fournissant un signal de supervision efficace pour l'apprentissage des classificateurs. Nous validons notre approche sur les ensembles de données EndoVis 2017 et 2018. Nous fournissons des résultats en utilisant des masques binaires obtenus soit par annotation manuelle, soit comme prédictions d'un modèle de segmentation binaire non supervisé. Cette dernière solution permet d'obtenir une approche de segmentation d'instance totalement exempte d'annotations par pixel, surpassant plusieurs approches de segmentation entièrement supervisées de l'état de l'art.

B.3.2 KI-BOT: une approche cinématique du goulot d'étranglement pour la régression de la pose d'instruments chirurgicaux flexibles directement à partir d'images

Comme dernière contribution, nous proposons une approche basée sur l'image et l'apprentissage profond pour l'estimation de la pose en 3D d'outils chirurgicaux flexibles. Les méthodes standard d'estimation de la pose des outils impliquent l'utilisation de capteurs externes, principalement des trackers électromagnétiques. Bien que précises, ces méthodes nécessitent des modifications indésirables des outils et éventuellement du flux de travail chirurgical. Le problème de l'estimation de la pose en 3D est formulé comme l'apprentissage d'un modèle de régresseur, afin d'estimer la pose des instruments chirurgicaux à partir des données d'image acquises par la caméra chirurgicale. Contrairement aux méthodes standard où la pose estimée est représentée par la position et l'orientation de l'effecteur, le régresseur est entraîné pour estimer directement les valeurs cinématiques des articulations des instruments. Les valeurs des articulations, combinées au modèle géométrique cinématique avancé des instruments, permettent de reconstruire leur forme 3D complète. En raison de l'absence d'un signal de supervision fiable (les valeurs cinématiques des articulations enregistrées par les systèmes robotiques sont généralement imprécises), nous exploitons une formulation d'auto-encodeur, basée sur l'idée de forcer la séparation de l'apparence d'une image de son contenu cinématique. L'architecture complète de l'apprentissage est présentée à la Figure B.6. Elle se compose de quatre modules:

- le régresseur cinématique (*kinematic regressor*), qui convertit l'image en un vecteur à faible dimension, chaque entrée correspondant à une valeur d'articulation spécifique. C'est le seul module qui sera conservé au moment de l'inférence;

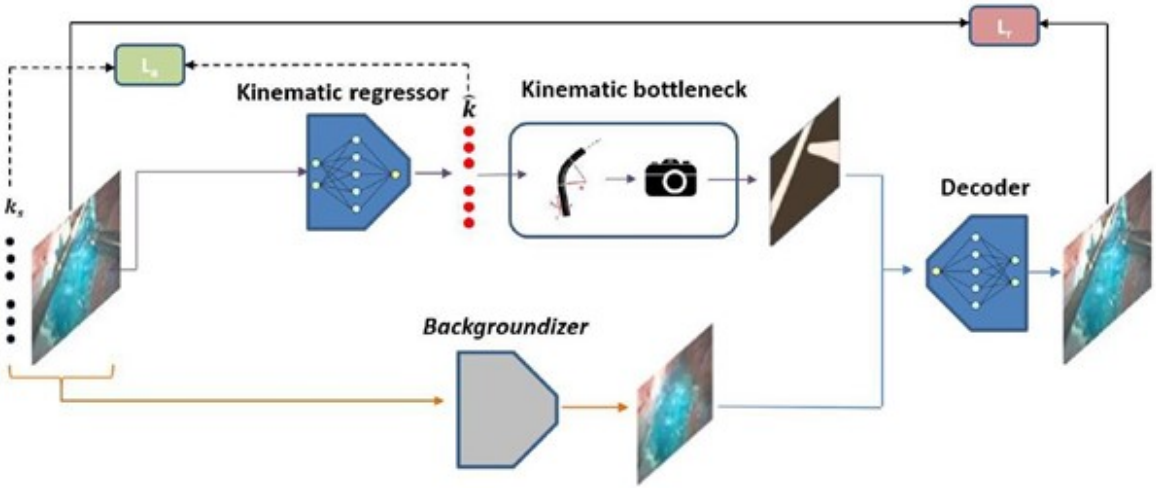

Figure B.6: Vue d'ensemble de l'architecture KI-BOT.

- le goulot d'étranglement cinématique (*kinematic bottleneck*), composé de deux sous-modules: le modèle géométrique cinématique direct des outils, qui met en correspondance les valeurs articulaires prédites avec la forme 3D correspondante des outils, et un moteur de rendu (modèle de caméra chirurgicale calibré et restériseur), qui reprojette la forme 3D dans l'espace image, afin d'obtenir un masque binaire, correspondant idéalement au masque de segmentation des outils dans l'image d'entrée ;

- le *backgroundizer*, qui masque les outils de l'image en exploitant la localisation approximative fournie par la cinématique enregistrée, et qui peint la région masquée afin de produire une version approximative de l'image d'entrée qui ne comporte que le fond;

- le décodeur (*decoder*), qui tente de reconstruire l'image d'entrée en combinant le masque binaire des outils et l'image d'arrière-plan.

Le goulot d'étranglement cinématique joue un double rôle pivot dans cette architecture. Tout d'abord, il permet de re-cartographier la tâche de l'espace 3D à l'espace 2D, grâce à l'opération de rendu: cette étape est nécessaire étant donné la nature différente de la pose estimée (3D) et de la vérité de terrain (l'image elle-même, 2D). Deuxièmement, il est nécessaire d'injecter des connaissances physiques du problème dans le processus d'apprentissage: en raison de la manière dont il traite les valeurs de sortie du régresseur, il lui donne la signification explicite de "cinématique", en associant à chaque valeur régressée une articulation spécifique. Le régresseur et le décodeur sont formés ensemble, en minimisant la distance entre les images d'entrée et les images reconstruites, sous la forme d'une perte perceptuelle. En outre, afin de tirer parti des informations fournies par la cinématique enregistrée, sans l'utiliser comme un signal de supervision fort, nous introduisons une "perte auxiliaire" d'erreur quadratique moyenne douce, comparant la cinématique prédite à celle enregistrée, et pénalisant le régresseur si la distance est supérieure à une certaine valeur de tolérance, qui tient compte de son incertitude. La validation a été effectuée à l'aide du système robotique STRAS, un robot endoscopique doté de deux bras robotisés. Les résultats qualitatifs sur les données in

vivo sont présentés à la Figure B.7. Les résultats quantitatifs ont montré une amélioration constante par rapport à la cinématique brute enregistrée, ce qui prouve que l'approche proposée est une solution valable pour l'estimation en temps réel de la pose 3D des instruments chirurgicaux basée sur l'image.

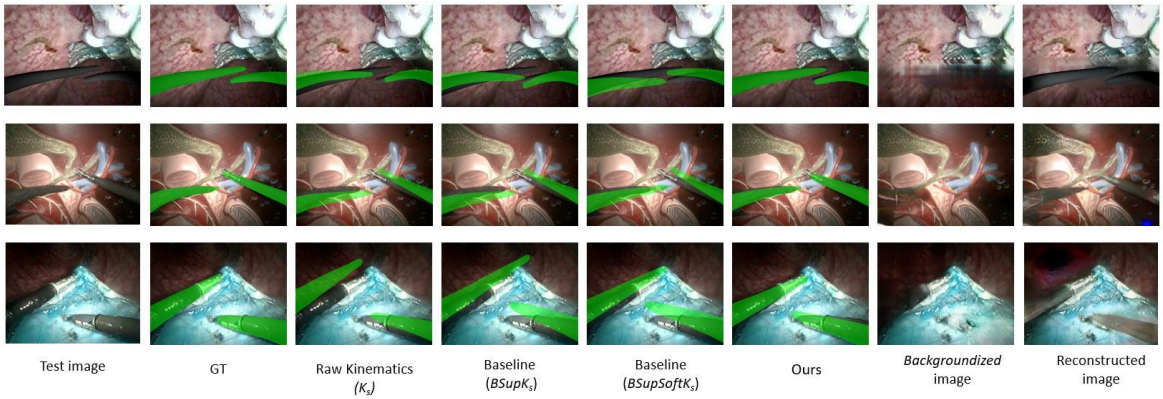

Figure B.7: Résultats qualitatifs sur l'ensemble de données STRAS in-vivo.

B.4 Conclusion

L'objectif des travaux présentés dans cette thèse était le développement de méthodes de localisation et d'identification d'instruments chirurgicaux, ne nécessitant pas de données annotées manuellement pour s'entraîner. Ceci a été obtenu en sélectionnant des sources d'information appropriées et en concevant des cadres capables d'assimiler ces informations, en extrayant des signaux de supervision efficaces pour l'entraînement des modèles d'apprentissage profond. Nous espérons que les approches que nous proposons faciliteront le développement de technologies d'assistance utiles pour améliorer la qualité des soins chirurgicaux.

Bibliography

[AAK+14] Stavros A Antoniou, George A Antoniou, Oliver O Koch, Rudolph Pointner, and Frank A Granderath. Meta-analysis of laparoscopic vs open cholecystectomy in elderly patients. *World Journal of Gastroenterology: WJG*, 20(46):17626, 2014.

[ACO+15] Max Allan, Ping-Lin Chang, Sébastien Ourselin, David J Hawkes, Ashwin Sridhar, John Kelly, and Danail Stoyanov. Image based surgical instrument pose estimation with multi-class labelling and optical flow. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18*, pages 331–338. Springer, 2015.

[AJB+17] Devansh Arpit, Stanislaw Jastrzkebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, pages 233–242. PMLR, 2017.

[AKB+20] Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al. 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190*, 2020.

[AMV+21] Deepak Alapatt, Pietro Mascagni, Armine Vardazaryan, Alain Garcia, Nariaki Okamoto, Didier Mutter, Jacques Marescaux, Guido Costamagna, Bernard Dallemagne, and Nicolas Padoy. Temporally constrained neural networks (tcnn): A framework for semi-supervised video semantic segmentation. *arXiv preprint arXiv:2112.13815*, 2021.

[Ans18] Maulana M Ansari. Optical illusions quintuple during laparoscopic total extraperitoneal preperitoneal (tepp) hernioplasty: A. *International Journal of Surgery*, 2(1):33–36, 2018.

[AOH+18] Max Allan, Sébastien Ourselin, David J Hawkes, John D Kelly, and Danail Stoyanov. 3-d pose estimation of articulated instruments in robotic minimally invasive surgery. *IEEE transactions on medical imaging*, 37(5):1204–1213, 2018.

[ASK+19] Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, et al. 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426*, 2019.

[AV06] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

[BAA+18] Sebastian Bodenstedt, Max Allan, Anthony Agustinos, Xiaofei Du, Luis Garcia-Peraza-Herrera, Hannes Kenngott, Thomas Kurmann, Beat Müller-Stich, Sebastien Ourselin, Daniil Pakhomov, et al. Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. *arXiv preprint arXiv:1805.02475*, 2018.

[BABG19] Federico Bolelli, Stefano Allegretti, Lorenzo Baraldi, and Costantino Grana. Spaghetti labeling: Directed acyclic graphs for block-based connected components labeling. *IEEE Transactions on Image Processing*, 29:1999–2012, 2019.

[BASJ17] David Bouget, Max Allan, Danail Stoyanov, and Pierre Jannin. Vision-based and marker-less surgical tool detection and tracking: a review of the literature. *Medical image analysis*, 35:633–654, 2017.

[BBO+15] David Bouget, Rodrigo Benenson, Mohamed Omran, Laurent Riffaud, Bernt Schiele, and Pierre Jannin. Detecting surgical tools by modelling local appearance and global shape. *IEEE transactions on medical imaging*, 34(12):2603–2617, 2015.

[BCF18] Mehdi Boroumand, Mo Chen, and Jessica Fridrich. Deep residual network for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 14(5):1181–1193, 2018.

[BDSF+16] Antonio Biondi, Carla Di Stefano, Francesco Ferrara, Angelo Bellia, Marco Vacante, and Luigi Piazza. Laparoscopic versus open appendectomy: a retrospective cohort study assessing outcomes and cost-effectiveness. *World Journal of Emergency Surgery*, 11(1):1–6, 2016.

[Bea22] Randy Bean. Moneyball 20 years later: A progress report on data and analytics in professional sports, 2022. [Online; accessed 11-March-2023].

[BGG15] Esther M Bonrath, Lauren E Gordon, and Teodor P Grantcharov. Characterising 'near miss' events in complex laparoscopic surgery through video analysis. *BMJ quality & safety*, 24(8):516–521, 2015.

[BNF+04] G Ross Baker, Peter G Norton, Virginia Flintoft, Régis Blais, Adalsteinn Brown, Jafna Cox, Ed Etchells, William A Ghali, Philip Hébert, Sumit R Majumdar, et al. The canadian adverse events study: the incidence of adverse events among hospital patients in canada. *Cmaj*, 170(11):1678–1686, 2004.

[Bor18] Lera Boroditsky. How language shapes the way we think. 2018.

[Cab16] Paolo Cabras. *3D pose estimation of continuously deformable instruments in robotic endoscopic surgery*. PhD thesis, Université de Strasbourg, 2016.

[CDRV09] Kaare Christensen, Gabriele Doblhammer, Roland Rau, and James W Vaupel. Ageing populations: the challenges ahead. *The lancet*, 374(9696):1196–1208, 2009.

[CES20] Emanuele Colleoni, Philip Edwards, and Danail Stoyanov. Synthetic and real inputs for tool segmentation in robotic surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 700–710. Springer, 2020.

[CFM+20] Nathan J Curtis, Jake D Foster, Danilo Miskovic, Chris SB Brown, Peter J Hewett, Sarah Abbott, George B Hanna, Andrew RL Stevenson, and Nader K Francis. Association of surgical skill assessment with clinical outcomes in cancer surgery. *JAMA surgery*, 155(7):590–598, 2020.

[CG15] Xinlei Chen and Abhinav Gupta. Webly supervised learning of convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1431–1439, 2015.

[CGD07] Magdalena K Chmarra, CA Grimbergen, and J Dankelman. Systems for tracking minimally invasive surgical instruments. *Minimally Invasive Therapy & Allied Technologies*, 16(6):328–340, 2007.

[CGD+21] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15334–15342, 2021.

[Cha22] B Chakravorti. Why ai failed to live up to its potential during the pandemic. *Harvard Business Review*, 2022.

[CK16] Kevin Cwach and Louis Kavoussi. Past, present, and future of laparoscopic renal surgery. *Investigative and clinical urology*, 57(Suppl 2):S110–S113, 2016.

[CKM05] Kevin Cleary, Audrey Kinsella, and Seong K Mun. Or 2020 workshop report: Operating room of the future. In *International Congress Series*, volume 1281, pages 832–838. Elsevier, 2005.

[CMB+18] Qinyu Chen, Katiuscha Merath, Fabio Bagante, Ozgur Akgul, Mary Dillhoff, Jordan Cloyd, and Timothy M Pawlik. A comparison of open and minimally invasive surgery for hepatic and pancreatic resections among the medicare population. *Journal of Gastrointestinal Surgery*, 22(12):2088–2096, 2018.

[CMS13] Kenneth Cukier and Viktor Mayer-Schoenberger. The rise of big data: How it's changing the way we think about the world. *Foreign Aff.*, 92:28, 2013.

[Coh13] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[CS21] Emanuele Colleoni and Danail Stoyanov. Robotic instrument segmentation with image-to-image translation. *IEEE Robotics and Automation Letters*, 6(2):935–942, 2021.

[CTM+21] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[CVMS20] François Chadebecq, Francisco Vasconcelos, Evangelos Mazomenos, and Danail Stoyanov. Computer vision in the surgical operating room. *Visceral Medicine*, 36(6):456–462, 2020.

[CZS17] Casey Chu, Andrey Zhmoginov, and Mark Sandler. Cyclegan, a master of steganography. *arXiv preprint arXiv:1712.02950*, 2017.

[dCRPR19] Cristian da Costa Rocha, Nicolas Padoy, and Benoit Rosa. Self-supervised surgical tool segmentation using kinematic information. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8720–8726. IEEE, 2019.

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[DDZZ+13] Antonio De Donno, Lucile Zorn, Philippe Zanne, Florent Nageotte, and Michel de Mathelin. Introducing stras: A new flexible robotic system for minimally invasive surgery. In *2013 IEEE International Conference on Robotics and Automation*, pages 1213–1220. IEEE, 2013.

[DGDM05] Christophe Doignon, Pierre Graebling, and Michel De Mathelin. Real-time segmentation of surgical instruments inside the abdominal cavity using a joint hue saturation color feature. *Real-Time Imaging*, 11(5-6):429–442, 2005.

[DHM+05] AJ Duffy, NJ Hogle, H McCarthy, JI Lew, A Egan, P Christos, and DL Fowler. Construct validity for the lapsim laparoscopic surgical simulator. *Surgical Endoscopy and Other Interventional Techniques*, 19(3):401–405, 2005.

[DNdM06] Christophe Doignon, Florent Nageotte, and Michel de Mathelin. The role of insertion points in the detection and positioning of instruments in laparoscopy for robotic tasks. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006: 9th International Conference, Copenhagen, Denmark, October 1-6, 2006. Proceedings, Part I 9*, pages 527–534. Springer, 2006.

[DWLU23] Hao Ding, Jie Ying Wu, Zhaoshuo Li, and Mathias Unberath. Rethinking causality-driven robot tool segmentation with temporal constraints. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–8, 2023.

[DZK+22] Hao Ding, Jintan Zhang, Peter Kazanzides, Jie Ying Wu, and Mathias Unberath. Carts: Causality-driven robot tool segmentation from vision and kinematics data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 387–398. Springer, 2022.

[EAAMA19] Omar Elharrouss, Noor Almaadeed, Somaya Al-Maadeed, and Younes Akbari. Image inpainting: A review. *Neural Processing Letters*, pages 1–22, 2019.

[FA+21] US Food, Drug Administration, et al. Good machine learning practice for medical device development: Guiding principles, 2021. [Online; accessed 06-March-2023].

[FCC+18] NK Francis, NJ Curtis, JA Conti, JD Foster, HJ Bonjer, and GB Hanna. Eaes classification of intraoperative adverse events in laparoscopic surgery. *Surgical Endoscopy*, 32:3822–3829, 2018.

[FMFV22] Marco Ferro, Alessandro Mirante, Fanny Ficuciello, and Marilena Vendittelli. A coppeliasim dynamic simulator for the da vinci research kit. *IEEE Robotics and Automation Letters*, 8(1):129–136, 2022.

[For19] Forbes Insights. The hospital will see you now, 2019. [Online; accessed 13-March-2023].

[GBSA20] Cristina González, Laura Bravo-Sánchez, and Pablo Arbelaez. Isinet: an instance-based approach for surgical instrument segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, pages 595–605. Springer, 2020.

[GCB04] Nizar Grira, Michel Crucianu, and Nozha Boujemaa. Unsupervised and semi-supervised clustering: a brief survey. *A review of machine learning techniques for processing multimedia content*, 1(2004):9–16, 2004.

[GDDM14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[GFK+19] Maria Grammatikopoulou, Evangello Flouty, Abdolrahim Kadkhodamohammadi, Gwenol'e Quellec, Andre Chow, Jean Nehme, Imanol Luengo, and Danail Stoyanov. Cadis: Cataract dataset for image segmentation. *arXiv preprint arXiv:1906.11586*, 2019.

[GIG] Gigenius. https://www.youtube.com/watch?v=AblniEzEgtY. [Online; accessed 11-March-2023].

[GKL+21] Carly R Garrow, Karl-Friedrich Kowalewski, Linhong Li, Martin Wagner, Mona W Schmidt, Sandy Engelhardt, Daniel A Hashimoto, Hannes G Kenngott, Sebastian Bodenstedt, Stefanie Speidel, et al. Machine learning for surgical phase recognition: a systematic review. *Annals of surgery*, 273(4):684–693, 2021.

[GPHFD+21] Luis C. Garcia-Peraza-Herrera, Lucas Fidon, Claudia D'Ettorre, Danail Stoyanov, Tom Vercauteren, and Sébastien Ourselin. Image compositing for segmentation of surgical tools without manual annotations. *IEEE Transactions on Medical Imaging*, 40(5):1450–1460, 2021.

[GPHLF+17] Luis C Garcia-Peraza-Herrera, Wenqi Li, Lucas Fidon, Caspar Gruijthuijsen, Alain Devreker, George Attilakos, Jan Deprest, Emmanuel Vander Poorten, Danail Stoyanov, Tom Vercauteren, et al. Toolnet: holistically-nested real-time segmentation of robotic surgical tools. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5717–5722. IEEE, 2017.

[GR18] Aniket Gadwe and Hongliang Ren. Real-time 6dof pose estimation of endoscopic instruments using printable markers. *IEEE Sensors Journal*, 19(6):2338–2346, 2018.

[Gra] Grabcad. https://grabcad.com/library/laparoscopic-forceps-head-1. [Online; accessed 06-March-2023].

[GSBW12] Berk Geveci, Will Schroeder, A Brown, and G Wilson. Vtk. *The architecture of open source applications*, 1:387–402, 2012.

[GZH+16] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016.

[HCS+14] Iliana J Harrysson, Jonathan Cook, Pramudith Sirimanna, Liane S Feldman, Ara Darzi, and Rajesh Aggarwal. Systematic review of learning curves for minimally invasive abdominal surgery: a review of the methodology of data collection, depiction of outcomes, and statistical analysis. *Annals of surgery*, 260(1):37–45, 2014.

[HFW+20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[HGDG17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[HKK+20] W-Y Hong, C-L Kao, Y-H Kuo, J-R Wang, W-L Chang, and C-S Shih. Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. *arXiv preprint arXiv:2012.12453*, 2020.

[HL19] SM Kamrul Hasan and Cristian A Linte. U-netplus: A modified encoder-decoder u-net architecture for semantic and instance segmentation of surgical instruments from laparoscopic images. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 7205–7211. IEEE, 2019.

[HSI+21] Cesare Hassan, Marco Spadaccini, Andrea Iannone, Roberta Maselli, Manol Jovani, Viveksandeep Thoguluva Chandrasekar, Giulio Antonelli, Honggang Yu, Miguel Areia, Mario Dinis-Ribeiro, et al. Performance of artificial intelligence in colonoscopy for adenoma and polyp detection: a systematic review and meta-analysis. *Gastrointestinal endoscopy*, 93(1):77–85, 2021.

[HYY+18] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[iR20] iData Research. Laparoscopic devices market size, share & covid-19 impact analysis, 2020. [Online; accessed 17-March-2023].

[IVLR21] Mobarakol Islam, VS Vibashan, Chwee Ming Lim, and Hongliang Ren. St-mtl: Spatio-temporal multitask learning model to predict scanpath while tracking instruments in robotic surgery. *Medical Image Analysis*, 67:101837, 2021.

[Jac15] Chevalier Jackson. *Peroral endoscopy and laryngeal surgery*. Laryngoscope Company, 1915.

[JBZ+20] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

[JCDH19a] Yueming Jin, Keyun Cheng, Qi Dou, and Pheng-Ann Heng. Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 440–448. Springer, 2019.

[JCDH19b] Yueming Jin, Keyun Cheng, Qi Dou, and Pheng-Ann Heng. Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video github repository, 2019.

[JCKK20] Ace St John, Ilaria Caturegli, Natalia S Kubicki, and Stephen M Kavic. The rise of minimally invasive surgery: 16 year analysis of the progressive replacement of open surgery with laparoscopy. *JSLS: Journal of the Society of Laparoscopic & Robotic Surgeons*, 24(4), 2020.

[JGBV20] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Self-supervised learning of interpretable keypoints from unlabelled videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8787–8797, 2020.

[JJLG20] James J Jung, Peter Jüni, Gerald Lebovic, and Teodor Grantcharov. First-year analysis of the operating room black box study. *Annals of surgery*, 271(1):122–127, 2020.

[JZL+18] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, pages 2304–2313. PMLR, 2018.

[KAN+21] Megha Kalia, Tajwar Abrar Aleef, Nassir Navab, Peter Black, and Septimiu E Salcudean. Co-generation and segmentation for generalized surgical instrument segmentation on unlabelled data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 403–412. Springer, 2021.

[Kid20] Robert Kidd. How one company is changing data and video analysis in soccer, 2020. [Online; accessed 12-March-2023].

[KJD+21] Xiaowen Kong, Yueming Jin, Qi Dou, Ziyi Wang, Zerui Wang, Bo Lu, Erbao Dong, Yun-Hui Liu, and Dong Sun. Accurate instance segmentation of surgical instruments in robotic surgery: Model refinement and cross-dataset evaluation. *International Journal of Computer Assisted Radiology and Surgery*, 16(9):1607–1614, 2021.

[KLO+20] Deborah R Kaye, Amy N Luckenbaugh, Mary Oerline, Brent K Hollenbeck, Lindsey A Herrel, Justin B Dimick, and John M Hollingsworth. Understanding the costs associated with surgical care delivery in the medicare population. *Annals of surgery*, 271(1):23, 2020.

[KMNA+21] Thomas Kurmann, Pablo Márquez-Neila, Max Allan, Sebastian Wolf, and Raphael Sznitman. Mask then classify: multi-instance segmentation for surgical instruments. *International journal of computer assisted radiology and surgery*, 16(7):1227–1236, 2021.

[KMR+23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[KTW+20] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[KUH18] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3907–3916, 2018.

[Lap20] LapSim. Lapsim. the proven simulator, 2020. [Online; accessed 06-March-2023].

[Las16] Lynn Lashbrook. Your basketball analytics skills can help you get discovered, 2016. [Online; accessed 12-March-2023].

[LC22] Xinye Li and Ding Chen. A survey on deep learning-based panoptic segmentation. *Digital Signal Processing*, 120:103283, 2022.

[LDH+17] András Légner, Michele Diana, Péter Halvax, Yu-Yin Liu, Lucile Zorn, Philippe Zanne, Florent Nageotte, Michel De Mathelin, Bernard Dallemagne, and Jacques Marescaux. Endoluminal surgical triangulation 2.0: A new flexible surgical robot. preliminary pre-clinical results with colonic submucosal dissection. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 13(3):e1819, 2017.

[LeC98] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[Lib] Library of alexandria. https://en.wikipedia.org/wiki/Library_of_Alexandria. [Online; accessed 01-March-2023].

[LKH+13] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg. Video segmentation by tracking many figure-ground segments. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2192–2199, 2013.

[LLGY23] Jingpei Lu, Fei Liu, Cédric Girerd, and Michael C Yip. Image-based pose estimation and shape reconstruction for robot manipulators and soft, continuum robots via differentiable rendering. *arXiv preprint arXiv:2302.14039*, 2023.

[Loh12] Steve Lohr. The age of big data. *New York Times*, 11(2012), 2012.

[LRR+17] Iro Laina, Nicola Rieke, Christian Rupprecht, Josué Page Vizcaíno, Abouzar Eslami, Federico Tombari, and Nassir Navab. Concurrent segmentation and localization for tracking of surgical instruments. In *International conference on medical image computing and computer-assisted intervention*, pages 664–672. Springer, 2017.

[LRS+18] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018.

[LWJ+20a] Daochang Liu, Yuhui Wei, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, and Ziyu Li. Unsupervised surgical instrument segmentation via anchor generation and semantic diffusion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 657–667. Springer, 2020.

[LWJ+20b] Daochang Liu, Yuhui Wei, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, and Ziyu Li. Unsupervised surgical instrument segmentation via anchor generation and semantic diffusion github repository, 2020.

[LWL+17] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017.

[LZK+21] Joël L Lavanchy, Joel Zindel, Kadir Kirtac, Isabell Twick, Enes Hosgor, Daniel Candinas, and Guido Beldi. Automation of surgical skill assessment using a three-stage machine learning algorithm. *Scientific reports*, 11(1):1–9, 2021.

[MAL+22] Pietro Mascagni, Deepak Alapatt, Giovanni Guglielmo Laracca, Ludovica Guerriero, Andrea Spota, Claudio Fiorillo, Armine Vardazaryan, Giuseppe Quero, Sergio Alfieri, Ludovica Baldari, et al. Multicentric validation of endodigest: a computer vision platform for video documentation of the critical view of safety in laparoscopic cholecystectomy. *Surgical Endoscopy*, 36(11):8379–8386, 2022.

[MAO+20] Sabarinath Mahadevan, Ali Athar, Aljoša Ošep, Sebastian Hennen, Laura Leal-Taixé, and Bastian Leibe. Making a case for 3d convolutions for object segmentation in videos. In *British Machine Vision Virtual Conference (BMVC)*, 2020.

[MAU+21] Pietro Mascagni, Deepak Alapatt, Takeshi Urade, Armine Vardazaryan, Didier Mutter, Jacques Marescaux, Guido Costamagna, Bernard Dallemagne, and Nicolas Padoy. A computer vision platform to automatically locate critical events in surgical videos: documenting safety in laparoscopic cholecystectomy. *Annals of surgery*, 274(1):e93–e95, 2021.

[MB18] Michael X Ma and Michael J Bourke. Endoscopic submucosal dissection in the west: current status and future directions. *Digestive Endoscopy*, 30(3):310–320, 2018.

[McG15] Geoff McGhee. Using maps and data vis to understand tennis, 2015. [Online; accessed 12-March-2023].

[Mes17] Bertalan Mesko. The role of artificial intelligence in precision medicine. *Expert Review of Precision Medicine and Drug Development*, 2(5):239–241, 2017.

[MGP+22] Sahar Mirzaee, Mahdieh Golzarand, Reza Parsaei, Karamollah Toolabi, and Alireza Amirbeigi. How accurate is the visual estimation of bowel length by endoscopic surgeons? *Frontiers in Surgery*, 9, 2022.

[MHEF+18] Lena Maier-Hein, Matthias Eisenmann, Carolin Feldmann, Hubertus Feussner, Germain Forestier, Stamatia Giannarou, Bernard Gibaud, Gregory D Hager, Makoto Hashizume, Darko Katic, et al. Surgical data science: A consensus perspective. *arXiv preprint arXiv:1806.03184*, 2018.

[MHES+22] Lena Maier-Hein, Matthias Eisenmann, Duygu Sarikaya, Keno März, Toby Collins, Anand Malpani, Johannes Fallert, Hubertus Feussner, Stamatia Giannarou, Pietro Mascagni, et al. Surgical data science–from concepts toward clinical translation. *Medical image analysis*, 76:102306, 2022.

[MHVS+17] Lena Maier-Hein, Swaroop S Vedula, Stefanie Speidel, Nassir Navab, Ron Kikinis, Adrian Park, Matthias Eisenmann, Hubertus Feussner, Germain Forestier, Stamatia Giannarou, et al. Surgical data science for next-generation interventions. *Nature Biomedical Engineering*, 1(9):691–696, 2017.

[MIH+16] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016.

[MISF20] Rocco Moccia, Cristina Iacono, Bruno Siciliano, and Fanny Ficuciello. Vision-based dynamic virtual fixtures for tools collision avoidance in robotic surgery. *IEEE Robotics and Automation Letters*, 5(2):1650–1655, 2020.

[MLF+19] Pietro Mascagni, Sun Gyo Lim, Claudio Fiorillo, Philippe Zanne, Florent Nageotte, Lucile Zorn, Silvana Perretta, Michel de Mathelin, Jacques Marescaux, and Bernard Dallemagne. Democratizing endoscopic submucosal dissection: single-operator fully robotic colorectal endoscopic submucosal dissection in a pig model. *Gastroenterology*, 156(6):1569–1571, 2019.

[MMC+21] Aldo Marzullo, Sara Moccia, Michele Catellani, Francesco Calimeri, and Elena De Momi. Towards realistic laparoscopic image generation using image-domain translation. *Computer Methods and Programs in Biomedicine*, 200:105834, 2021.

[MMIW10] Eric Muñoz, William Muñoz III, and Leslie Wise. National and surgical health care expenditures, 2005–2025. *Annals of surgery*, 251(2):195–200, 2010.

[MRD+07] CB Morgenthal, WO Richards, Brian J Dunkin, KA Forde, G Vitale, E Lin, and SAGES Flexible Endoscopy Committee. The role of the surgeon in the evolution of flexible endoscopy. *Surgical endoscopy*, 21:838–853, 2007.

[Nak17] Don K Nakayama. The minimally invasive operations that transformed surgery. *American College of Surgeons*, 2017.

[NGWS18] Dong Nie, Yaozong Gao, Li Wang, and Dinggang Shen. Asdnet: Attention based semi-supervised deep networks for medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 370–378. Springer, 2018.

[NMB+19] Dmitri Nepogodiev, Janet Martin, Bruce Biccard, Alex Makupe, Aneel Bhangu, Adesoji Ademuyiwa, Adewale Oluseye Adisa, Maria-Lorena Aguilera, Sohini Chakrabortee, J Edward Fitzgerald, et al. Global burden of postoperative death. *The Lancet*, 393(10170):401, 2019.

[NMMP19] Chinedu Innocent Nwoye, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Weakly supervised convolutional lstm approach for tool tracking in laparoscopic videos. *International journal of computer assisted radiology and surgery*, 14:1059–1067, 2019.

[NMN+20] Tam Nguyen, C Mummadi, T Ngo, L Beggel, and Thomas Brox. Self: learning to filter noisy labels with self-ensembling. In *International Conference on Learning Representations (ICLR)*, 2020.

[NP22] Chinedu Innocent Nwoye and Nicolas Padoy. Data splits and metrics for method benchmarking on surgical action triplet datasets. *arXiv preprint arXiv:2204.05235*, 2022.

[NYG+22] Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433, 2022.

[OMB13] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1187–1200, 2013.

[PBA+12] Nicolas Padoy, Tobias Blum, Seyed-Ahmad Ahmadi, Hubertus Feussner, Marie-Odile Berger, and Nassir Navab. Statistical modeling and recognition of surgical workflow. *Medical image analysis*, 16(3):632–641, 2012.

[PFR+19] Micha Pfeiffer, Isabel Funke, Maria R Robu, Sebastian Bodenstedt, Leon Strenger, Sandy Engelhardt, Tobias Roß, Matthew J Clarkson, Kurinchi Gurusamy, Brian R Davidson, et al. Generating large labeled data sets for laparoscopic image processing tasks using unpaired image-to-image translation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22*, pages 119–127. Springer, 2019.

[PPA+19] Daniil Pakhomov, Vittal Premachandran, Max Allan, Mahdi Azizian, and Nassir Navab. Deep residual learning for instrument segmentation in robotic surgery. In *International Workshop on Machine Learning in Medical Imaging*, pages 566–573. Springer, 2019.

[PPTM+16] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016.

[Pre13] Gil Press. A very short history of big data, 2013. [Online; accessed 16-March-2023].

[Pri21] Steve Price. How premier league stars can use statistics to boost their careers, 2021. [Online; accessed 12-March-2023].

[PSN20] Daniil Pakhomov, Wei Shen, and Nassir Navab. Towards unsupervised learning for instrument segmentation in robotic surgery with cycle-consistent adversarial networks. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8499–8504. IEEE, 2020.

[PVES21] Krittin Pachtrachai, Francisco Vasconcelos, Philip Edwards, and Danail Stoyanov. Learning to calibrate-estimating the hand-eye transformation without calibration objects. *IEEE Robotics and Automation Letters*, 6(4):7309–7316, 2021.

[PVH09] Zachary Pezzementi, Sandrine Voros, and Gregory D Hager. Articulated object tracking by rendering consistent appearance parts. In *2009 IEEE International Conference on Robotics and Automation*, pages 3940–3947. IEEE, 2009.

[PWA+19] Kenneth A Philbrick, Alexander D Weston, Zeynettin Akkus, Timothy L Kline, Panagiotis Korfiatis, Tomas Sakinis, Petro Kostandy, Arunnit Boonrod, Atefeh Zeinoddini, Naoki Takahashi, et al. Ril-contour: a medical imaging dataset annotation tool for and with deep learning. *Journal of digital imaging*, 32:571–581, 2019.

[RDM89] Harry Reich, JOHN DeCAPRIO, and FRAN McGLYNN. Laparoscopic hysterectomy. *Journal of Gynecologic Surgery*, 5(2):213–216, 1989.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[Ric16] Ruthann Richter. Stanford-led study underscores huge gap between rich, poor in global surgery, 2016. [Online; accessed 13-March-2023].

[RJ01] Walker Reynolds Jr. The first laparoscopic cholecystectomy. *JSLS: Journal of the Society of Laparoendoscopic Surgeons*, 5(1):89, 2001.

[RRF+20] Tobias Ross, Annika Reinke, Peter M Full, Martin Wagner, Hannes Kenngott, Martin Apitz, Hellena Hempe, Diana Mindroc Filimon, Patrick Scholz, Thuy Nuong Tran, et al. Robust medical instrument segmentation challenge 2019. *arXiv preprint arXiv:2003.10299*, 2020.

[RS21] Khalid Raza and Nripendra K Singh. A tour of unsupervised deep learning for medical image analysis. *Current Medical Imaging*, 17(9):1059–1077, 2021.

[RSA+22] Sanat Ramesh, Vinkle Srivastav, Deepak Alapatt, Tong Yu, Aditya Murali, Luca Sestini, Chinedu Innocent Nwoye, Idris Hamoud, Antoine Fleurentin, Georgios Exarchakis, et al. Dissecting self-supervised learning methods for surgical computer vision. *arXiv preprint arXiv:2207.00449*, 2022.

[RVBS17] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.

[RZV+18] Tobias Ross, David Zimmerer, Anant Vemuri, Fabian Isensee, Manuel Wiesenfarth, Sebastian Bodenstedt, Fabian Both, Philip Kessler, Martin Wagner, Beat Müller, et al. Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. *International journal of computer assisted radiology and surgery*, 13(6):925–933, 2018.

[SAG14] SAGES. Society of american gastrointestinal and endoscopic surgeons (sages). https://www.sages.org/image-tag/coaxial-alignment/, 2014. [Online; accessed 06-March-2023].

[Sat03] Richard M Satava. Disruptive visions: the operating room of the future. *Surgical endoscopy*, 17(1):104–107, 2003.

[Saw09] Shlomo S Sawilowsky. New effect size rules of thumb. *Journal of modern applied statistical methods*, 8(2):26, 2009.

[SBAM15] Mark G Shrime, Stephen W Bickler, Blake C Alkire, and Charlie Mock. Global burden of surgical disease: an estimation from the provider perspective. *The Lancet Global Health*, 3:S8–S9, 2015.

[SBF14] Raphael Sznitman, Carlos Becker, and Pascal Fua. Fast part-based classification for instrument detection in minimally invasive surgery. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part II 17*, pages 692–699. Springer, 2014.

[SC00] GEI Shallaly and A Cuschieri. Nature, aetiology and outcome of bile duct injuries after laparoscopic cholecystectomy. *HPB*, 2(1):3–12, 2000.

[Sch16] Klaus Schwab. The fourth industrial revolution: what it means, how to respond, 2016. [Online; accessed 06-March-2023].

[SFV+10] Gideon Sroka, Liane S Feldman, Melina C Vassiliou, Pepa A Kaneva, Raad Fayez, and Gerald M Fried. Fundamentals of laparoscopic surgery simulator training to proficiency improves laparoscopic performance in the operating room—a randomized controlled trial. *The American journal of surgery*, 199(1):115–120, 2010.

[SHK+20] Jonah J Stulberg, Reiping Huang, Lindsey Kreutzer, Kristen Ban, Bradley J Champagne, Scott R Steele, Julie K Johnson, Jane L Holl, Caprice C Greenberg, and Karl Y Bilimoria. Association between surgeon technical skills and patient outcomes. *JAMA surgery*, 155(10):960–968, 2020.

[SHW+10] Lygia Stewart, John G Hunter, Alberto Wetter, Brian Chin, and Lawrence W Way. Operative reports: form and function. *Archives of Surgery*, 145(9):865–871, 2010.

[SJDH21] Xueying Shi, Yueming Jin, Qi Dou, and Pheng-Ann Heng. Semi-supervised learning with progressive unlabeled data excavation for label-efficient surgical workflow recognition. *Medical Image Analysis*, 73:102158, 2021.

[SLQ+16] Chaoyang Shi, Xiongbiao Luo, Peng Qi, Tianliang Li, Shuang Song, Zoran Najdovski, Toshio Fukuda, and Hongliang Ren. Shape sensing techniques for continuum robots in minimally invasive surgery: A survey. *IEEE Transactions on Biomedical Engineering*, 64(8):1665–1678, 2016.

[SMZ21] Manish Sahu, Anirban Mukhopadhyay, and Stefan Zachow. Simulation-to-real domain adaptation with teacher–student learning for endoscopic instrument segmentation. *International journal of computer assisted radiology and surgery*, 16(5):849–859, 2021.

[SRDM+21] Luca Sestini, Benoit Rosa, Elena De Momi, Giancarlo Ferrigno, and Nicolas Padoy. A kinematic bottleneck approach for pose regression of flexible surgical instruments directly from images. *IEEE Robotics and Automation Letters*, 6(2):2938–2945, 2021.

[SRDM+23] Luca Sestini, Benoit Rosa, Elena De Momi, Giancarlo Ferrigno, and Nicolas Padoy. Fun-sis: A fully unsupervised approach for surgical instrument segmentation. *Medical Image Analysis*, page 102751, 2023.

[Sri21] Vinkle Kumar Srivastav. *Unsupervised domain adaptation approaches for person localization in the operating rooms*. PhD thesis, Université de Strasbourg, 2021.

[SRKI18a] Alexey A Shvets, Alexander Rakhlin, Alexandr A Kalinin, and Vladimir I Iglovikov. Automatic instrument segmentation in robot-assisted surgery using deep learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 624–628. IEEE, 2018.

[SRKI18b] Alexey A Shvets, Alexander Rakhlin, Alexandr A Kalinin, and Vladimir I Iglovikov. Automatic instrument segmentation in robot-assisted surgery using deep learning github repository, 2018.

[SSMZ20] Manish Sahu, Ronja Strömsdörfer, Anirban Mukhopadhyay, and Stefan Zachow. Endo-sim2real: Consistency learning-based domain adaptation for instrument segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 784–794. Springer, 2020.

[SW07] Lygia Stewart and Lawrence W Way. The prevention of laparoscopic bile duct injuries: an analysis of 300 cases of from a human factors and cognitive psychology perspective. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 51, pages 617–620. SAGE Publications Sage CA: Los Angeles, CA, 2007.

[SYLK18] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018.

[SZ14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[TBB17] Eduardo M Targarona, Andrea Balla, and Gabriela Batista. Big data and surgery: The digital revolution continues. *Cirugia Espanola*, 96(5):247–249, 2017.

[TD20] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020.

[TKM+20] Yuichi Takayama, Yuji Kaneoka, Atsuyuki Maeda, Takamasa Takahashi, and Masahito Uji. Laparoscopic transabdominal preperitoneal repair versus open mesh plug repair for bilateral primary inguinal hernia. *Annals of gastroenterological surgery*, 4(2):156–162, 2020.

[TPPV21] Leonardo Tanzi, Pietro Piazzolla, Francesco Porpiglia, and Enrico Vezzetti. Real-time deep learning semantic segmentation during intra-operative surgery for 3d augmented reality assistance. *International Journal of Computer Assisted Radiology and Surgery*, 16(9):1435–1445, 2021.

[TSM+16] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016.

[VDDP18] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018.

[VdMH08] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[VLC06] Sandrine Voros, Jean-Alexandre Long, and Philippe Cinquin. Automatic localization of laparoscopic instruments for the visual servoing of an endoscopic camera holder. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006: 9th International Conference, Copenhagen, Denmark, October 1-6, 2006. Proceedings, Part I 9*, pages 535–542. Springer, 2006.

[VMMP18] Armine Vardazaryan, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Weakly-supervised learning for tool localization in laparoscopic videos. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 169–179. Springer, 2018.

[WAH97] Guo-Qing Wei, Klaus Arbter, and Gerd Hirzinger. Automatic tracking of laparoscopic instruments by color coding. In *CVRMed-MRCAS'97: First Joint Conference Computer Vision, Virtual Reality and Robotics in Medicine and Medical Robotics and Computer-Assisted Surgery Grenoble, France, March 19–22, 1997 Proceedings*, pages 357–366. Springer, 1997.

[Wel19] Bryce Welker. Data science, the fourth industrial revolution and the future of entrepreneurship, 2019. [Online; accessed 06-March-2023].

[WHM+15] Thomas G Weiser, Alex B Haynes, George Molina, Stuart R Lipsitz, Micaela M Esquivel, Tarsicio Uribe-Leitz, Rui Fu, Tej Azad, Tiffany E Chao, William R Berry, et al. Estimate of the global volume of surgery in 2012: an assessment supporting improved health outcomes. *The Lancet*, 385:S11, 2015.

[WIJ10] Robert J Webster III and Bryan A Jones. Design and kinematic modeling of constant curvature continuum robots: A review. *The International Journal of Robotics Research*, 29(13):1661–1683, 2010.

[WLFD22] Yuehao Wang, Yonghao Long, Siu Hin Fan, and Qi Dou. Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 431–441. Springer, 2022.

[WMB+18] Martin Wagner, Benjamin Friedrich Berthold Mayer, Sebastian Bodenstedt, Katherine Stemmer, Arash Fereydooni, Stefanie Speidel, Rüdiger Dillmann, Felix Nickel, Lars Fischer, and Hannes Götz Kenngott. Computer-assisted 3d bowel length measurement for quantitative laparoscopy. *Surgical Endoscopy*, 32:4052–4061, 2018.

[WMC+19] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 322–330, 2019.

[Wor] World Laparoscopy Hospital. https://www.laparoscopyhospital.com/. [Online; accessed 10-February-2022].

[WS20] Yu-Shin Wang and Kai-Tai Song. Image-based pose estimation and tracking of surgical instruments in minimally invasive surgery. In *2020 International Automatic Control Conference (CACS)*, pages 1–6. IEEE, 2020.

[WSG+03] Lawrence W Way, Lygia Stewart, Walter Gantert, Kingsway Liu, Crystine M Lee, Karen Whang, and John G Hunter. Causes and prevention of laparoscopic bile duct injuries: analysis of 252 cases from a human factors and cognitive psychology perspective. *Annals of surgery*, 237(4):460, 2003.

[WST+13] Meagan E Wiebe, Lakhbir Sandhu, Julie L Takata, Erin D Kennedy, Nancy N Baxter, Anna R Gagliardi, David R Urbach, and Alice C Wei. Quality of narrative operative reports in pancreatic surgery. *Canadian Journal of Surgery*, 56(5):E121, 2013.

[WSYP17] Wenguan Wang, Jianbing Shen, Ruigang Yang, and Fatih Porikli. Saliency-aware video object segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 40(1):20–33, 2017.

[XLL+22] Yao Xue, Siming Liu, Yonghui Li, Ping Wang, and Xueming Qian. A new weakly supervised strategy for surgical tool detection. *Knowledge-Based Systems*, 239:107860, 2022.

[XWY+21] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

[XXY+15] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015.

[YDDM18] Jie Yang, Thomas Drake, Andreas Damianou, and Yoelle Maarek. Leveraging crowdsourcing data for deep active learning an application: Learning intents in alexa. In *Proceedings of the 2018 World Wide Web Conference*, pages 23–32, 2018.

[YLL+21] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7177–7188, 2021.

[YLSS19a] Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, and Stefano Soatto. Unsupervised moving object detection via contextual information separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 879–888, 2019.

[YLSS19b] Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, and Stefano Soatto. Unsupervised moving object detection via contextual information separation github repository, 2019.

[YSL22] Zixin Yang, Richard Simon, and Cristian Linte. A weakly supervised learning approach for surgical instrument segmentation from laparoscopic video sequences. In *Medical Imaging 2022: image-Guided Procedures, Robotic Interventions, and Modeling*, volume 12034, pages 412–417. SPIE, 2022.

[YWL+19] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 605–613. Springer, 2019.

[ZBL+23] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, and Xia Hu. Data-centric ai: Perspectives and challenges. *arXiv preprint arXiv:2301.04819*, 2023.

[ZdBdK+11] Marieke Zegers, Martine C de Bruijne, Bertus de Keizer, Hanneke Merten, Peter P Groenewegen, Gerrit van der Wal, and Cordula Wagner. The incidence, root-causes, and outcomes of adverse events in surgical units: implication for potential prevention strategies. *Patient safety in surgery*, 5(1):1–11, 2011.

[Žel23] Jakub Železnỳ. Why competition is bad for science. *Nature Physics*, pages 1–1, 2023.

[ZJG+20] Zixu Zhao, Yueming Jin, Xiaojie Gao, Qi Dou, and Pheng-Ann Heng. Learning motion flows for semi-supervised instrument segmentation from robotic surgical video. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 679–689. Springer, 2020.

[ZJH22] Zixu Zhao, Yueming Jin, and Pheng-Ann Heng. Trasetr: track-to-segment transformer with contrastive query for instance-level instrument segmentation in robotic surgery. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 11186–11193. IEEE, 2022.

[ZJY+21] Di Zhang, Fan Jiang, Rui Yin, Ge-Ge Wu, Qi Wei, Xin-Wu Cui, Shu-E Zeng, Xue-Jun Ni, and Christoph F Dietrich. A review of the role of the s-detect computer-aided diagnostic ultrasound system in the evaluation of benign and malignant breast and thyroid masses. *Medical science monitor: international medical journal of experimental and clinical research*, 27:e931957–1, 2021.

[ZLZ+19] Marco A Zenati, Kay B Leissner, Suzana Zorca, Lauren Kennedy-Metz, Steven J Yule, and Roger D Dias. First reported use of team cognitive workload for root cause analysis in cardiac surgery. In *Seminars in thoracic and cardiovascular surgery*, volume 31, pages 394–396. Elsevier, 2019.

[ZMSO20] Zohreh Zakeri, Neil Mansfield, Caroline Sunderland, and Ahmet Omurtag. Physiological correlates of cognitive load in laparoscopic surgery. *Scientific reports*, 10(1):1–13, 2020.

[ZNZ+17] Lucile Zorn, Florent Nageotte, Philippe Zanne, Andras Legner, Bernard Dallemagne, Jacques Marescaux, and Michel de Mathelin. A novel telemanipulated robotic assistant for surgical endoscopy: preclinical application to esd. *IEEE Transactions on Biomedical Engineering*, 65(4):797–808, 2017.

[ZPIE17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[ZRCM+21] Zhongkai Zhang, Benoît Rosa, Oscar Caravaca-Mora, Philippe Zanne, Michalina J Gora, and Florent Nageotte. Image-guided control of an endoscopic robot for oct path scanning. *IEEE Robotics and Automation Letters*, 6(3):5881–5888, 2021.

[ZS18] Zhilu Zhang and Mert R Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *32nd Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

Luca SESTINI

Methods for Learning Surgical Instrument Segmentation from Unlabelled Datasets using Prior Knowledge

Résumé

Les vidéos chirurgicales constituent une riche source d'informations pour l'étude et l'amélioration de la chirurgie mini-invasive. L'identification et la localisation des instruments chirurgicaux à partir de ces vidéos est une étape cruciale pour le développement d'applications telles que l'évaluation automatique des compétences chirurgicales et l'aide à la décision en temps réel. Les approches de pointe pour une telle tâche reposent sur l'apprentissage entièrement supervisé de modèles d'apprentissage profond, nécessitant des données annotées manuellement, difficiles à collecter à grande échelle. Cette thèse propose des méthodes d'apprentissage profond pour la localisation et l'identification des instruments, qui peuvent être entraînées sur des ensembles de données totalement non étiquetées. Les connaissances générales sur les instruments chirurgicaux - moins chères et plus faciles à obtenir que les annotations manuelles - sont incorporées dans les architectures d'apprentissage pour fabriquer des signaux de supervision efficaces. Les approches proposées s'appuient sur de nouvelles méthodes d'apprentissage non supervisé, d'apprentissage de représentation auto-supervisé et d'apprentissage à partir d'étiquettes bruitées, toutes conçues pour exploiter efficacement ces connaissances préalables et complémentaires. Nous espérons que les approches que nous proposons pourront faciliter le développement de technologies d'assistance permettant d'améliorer la qualité des soins chirurgicaux.

Mots-clés : Apprentissage profond – Vision par ordinateur – Apprentissage auto-supervisé – Endoscopie – Segmentation d'instruments chirurgicaux

Résumé en anglais

Surgical videos offer a rich source of information for studying and improving minimally invasive surgery. Identifying and localising surgical instruments from these videos is a crucial step for the development of valuable applications like automatic surgical skill assessment and real-time decision support. State-of-the-art approaches for such task rely on fully-supervised training of deep learning models, requiring manually annotated data, hard to collect at a large scale. This thesis proposes deep learning methods for instrument localisation and identification which can be trained on completely unlabelled datasets. General knowledge about surgical instruments - cheaper and more easily obtained than manual annotations - is incorporated in the training architectures to fabricate effective supervision signals. The proposed approaches leverage novel methods for unsupervised learning, self-supervised representation learning, and learning from noisy labels, all designed to effectively mine such prior and complementary knowledge. We hope our proposed approaches can facilitate the development of valuable assistive technologies to improve the quality of surgical care.

Keywords: Deep learning – Computer vision – Self-supervised learning – Endoscopy – Surgical Instrument Segmentation