Contexte global avec des Transformers pour la segmentation d'images médicales 3D

Loic Themyr

HESAM UNIVERSITÉ

le cnam

ÉCOLE DOCTORALE Sciences des Métiers de l'Ingénieur
Centre d'études et de recherche en informatique et communications

THÈSE

présentée par : **Loic THEMYR**
soutenue le : **21 Décembre 2023**

pour obtenir le grade de : **Docteur d'HESAM Université**

préparée au : **Conservatoire national des arts et métiers**
Discipline : **Sciences et technologies de l'information et de la communication**
Spécialité : **Informatique**

Global context with transformers in 3D Medical Image Segmentation

THÈSE DIRIGÉE PAR :
M. THOME Nicolas Professeur des unniversités, CNAM/Sorbonne Université

Jury

Mme Caroline PETITJEAN	Professeure, LITIS, Université de Rouen Normandie	Rapporteur
M. Olivier BERNARD	Professeur, CREATIS, University de Lyon	Rapporteur
M. Christian DEROSIERS	Professeur, Université du Québec	Examinateur
M. Daniel GEORGE	Professeur, Université de Strasbourg	Président
M. Nicolas THOME	Professeur, CNAM, Sorbonne Université	Directeur
M. Clément RAMBOUR	Maître de conférences, CNAM	Co-encadrant
M. Toby COLLINS	Directeur de recherche, IRCAD Strasbourg	Co-encadrant
M. Alexandre HOSTETTLER	Directeur scientifique, IRCAD Strasbourg	Invité

THÈSE

Remerciements

Il m'est difficile d'exprimer individuellement mes remerciements envers toutes les personnes qui m'ont soutenu et qui ont rendu possible l'achèvement de cette thèse.

En premier lieu, je tiens à adresser mes sincères remerciements au Professeur Nicolas THOME, qui m'a chaleureusement accueilli au sein de son équipe de recherche au laboratoire CEDRIC du CNAM. Sa confiance et son accompagnement tout au long de ces trois années ont été inestimables. Il s'est investi même dans les moments les plus exigeants, et je lui en suis profondément reconnaissant. Je souhaite également exprimer ma gratitude envers Clément RAMBOUR, qui m'a encadré avec une implication remarquable et m'a guidé à travers les défis rencontrés durant ces trois années.

Je tiens à remercier tout particulièrement Toby COLLINS et Alexandre HOSTETTLER, sans lesquels cette thèse n'aurait pas pu voir le jour. Leur confiance en mon travail, du début à la fin, a été un pilier essentiel. Je leur suis également reconnaissant pour l'expertise précieuse qu'ils ont partagée avec moi. Mes remerciements vont également à William NDZIMBONG, avec qui j'ai partagé un projet récemment.

Mes camarades du CNAM, Elias RAMZI, Perla DOUBINSKY, Yannis KARMIM, Marc LAFON, Laura CALEM, ont été mes compagnons de route depuis le début de cette aventure. Nous avons partagé de nombreux moments mémorables, et je les en remercie chaleureusement. Olivier PETIT et Denis COQUENET, avec qui j'ai collaboré sur un projet, méritent également mes remerciements.

Je remercie également la Professeure Caroline PETITJEAN et le Professeur Olivier BERNARD pour leur participation aux comités de suivi. Leurs conseils avisés ont joué un rôle fondamental dans la réalisation de cette thèse.

Bien évidemment, je souhaite exprimer ma reconnaissance à tous les membres du jury pour l'intérêt qu'ils ont porté à mes travaux, ainsi que pour avoir accepté de lire ma thèse et d'assister à la soutenance qui marque la conclusion de ces quatre années de recherche.

Enfin, je tiens à remercier du fond du cœur ma famille et mes amis. Leur soutien indéfectible et leur présence inestimable dans les moments les plus difficiles m'ont permis d'aller jusqu'au bout de cette aventure. Je tiens particulièrement à remercier Wilfrid THEMYR et Régine THEMYR pour leur

soutien constant malgré la distance, Eva THEMYR pour son aide précieuse au quotidien pendant cette dernière phase de ma thèse, ainsi que Maé THEMYR, Noan THEMYR, mes amis de Rêve Jeune et MC.

Résumé

Cette thèse aborde la tâche complexe de la segmentation d'images à haute dimension, en particulier les images médicales en 3D. La haute dimensionnalité inhérente de ces images pose un obstacle à une segmentation efficace. Les modèles d'apprentissage profond nécessitent un contexte global et des informations fines pour segmenter avec précision des structures complexes. Ces exigences ne sont, cependant, pas satisfaites par les modèles classiques tels que UNet, qui restent limités par la taille de leur champs réceptifs ainsi que des contraintes mémoire sur la taille des entrées de ces modèles. Pour lever ces limitations, nous avons étudié dans cette thèse les modèles d'attention Transformers, qui sont utilisés pour leur capacité à capturer des interactions à longue portée. Dans ce travail, nous introduisons des modules Transformers enrichis de représentations globales améliorant ainsi leur efficacité dans la modélisation d'interactions à longue portée et très longue portée dans des volumes de hautes dimensions.

Dans un premier temps, nous nous concentrons sur la segmentation d'images médicales en 2D et présentons le modèle U-Transformer. Ce modèle novateur intègre les architectures Transformer dans la segmentation d'imagerie médicale pour surmonter les limitations des modèles classiques UNet avec des champs réceptifs restreints. U-Transformer intègre des mécanismes d'auto-attention dans l'encodeur et plusieurs couches de mécanismes d'attention croisée dans le décodeur. En combinant les réseaux neuronaux convolutionnels avec les Transformers, U-Transformer atteint des performances de pointe sur deux ensembles de données différents, dépassant les résultats établient par nnUNet.

Bien que prometteuse, l'architecture U-Transformer ne modèle les interactions qu'au plus profond du modèle, ce qui entraîne une résolution dégradée. Une solution serait d'avoir des Transformers à chaque niveau de résolution, mais les besoin requis en termes de calcul et de mémoire sont alors irréalistes. Nous avons introduit GLAM pour résoudre ces problèmes. GLAM est un module conçu pour une intégration transparente dans les modèles Window Transformer. GLAM aborde les limitations auxquelles sont confrontés les Window Transformers, qui ont du mal à capturer des interactions à longue distance dans des cartes de caractéristiques de haute résolution. En utilisant des jetons globaux et des modules Transformer spécifiques, GLAM facilite la propagation de l'information entre les fenêtres, permettant aux fenêtres interconnectées de capturer des informations à longue portée.

GLAM surpasse les méthodes traditionnelles sur des ensembles de données de segmentation de scènes réelles et un ensemble de données de segmentation d'images médicales en 3D.

Dans le contexte de la segmentation d'images médicales 3D, GLAM est entraîné sur des patchs extraits des images d'origine et ne modélise pas les informations au-delà de ces régions, ce qui est important pour la segmentation de structures complexes nécessitant une information sur la strucutre anatomique imagée. Pour remédier à cela, nous avons proposé d'adapter le mécanisme de jeton global pour modéliser ces informations perdues. Ainsi, nous proposons FINE et LORI, qui permettent la modélisation d'interactions à longue portée et hors de portée. FINE et LORI sont des modules polyvalent pour diverses méthodes d'apprentissage profond. Les expériences menées sur trois ensembles de données de segmentation d'images médicales en 3D, comprenant des scans CT et des données d'échographie, démontrent de manière constante la supériorité de LORI par rapport aux méthodes classiques, soulignant sa robustesse et l'importance de la modélisation du contexte.

Mots-clés : Ségmentation, Transformer, Vision Artificielle, Apprentissage profond, Imagerie Médicale

Abstract

This thesis addresses the challenging task of segmenting high-dimensional images, specifically 3D medical images. The inherent high dimensionality of these images poses an obstacle to effective segmentation. We established that deep learning models require a global context and fine grain information to accurately segment complex structures. These requirements are not satisfied together by classical models like UNet which often struggle due to limited receptive fields and restricted input region sizes. To tackle these limitations, we studied in this thesis Transformers attention models which are employed for their capacity to capture long-range interactions. In this work, we introduce Transformer modules enriched with global tokens to enhance their effectiveness in modeling long and out-of-range interactions in high-dimensional images, such as 3D medical images.

First, we focus on 2D medical image segmentation and introduce the U-Transformer model. This pioneering model incorporates Transformer architectures into medical imaging segmentation to overcome the limitations of classical UNet models with restricted receptive fields. The U-Transformer integrates self-attention mechanisms in the encoder and multiple layers of cross-attention mechanisms in the decoder. By combining Convolutional Neural Networks with Transformers, the U-Transformer achieves state-of-the-art performance on two diverse datasets, surpassing the established state-of-the-art nnUNet baseline.

U-Transformer is interesting but only model interactions in the bottleneck, thus have a degraded resolution. The solution would be to have Transformers at each resolution level, but this implies computational and memory limitations. We introduced GLAM which tackles these issues. GLAM is a module designed for seamless integration into windowed Transformer models. GLAM addresses the limitations faced by prior windowed Transformers, which struggle to capture long-range interactions in high-resolution feature maps. Leveraging global tokens and specific Transformer modules, GLAM facilitates information propagation between windows, enabling interconnected windows to capture long-range information. GLAM outperforms traditional methods on real-life scene segmentation datasets and a 3D medical image segmentation dataset.

In the context of 3D medical image segmentation, GLAM is trained on cropped patches of the

full size images and doesn't model information beyond this cropped regions which is important for complex structures which requires more information to be segmented. To address this, we proposed to adapt the global token mechanism to model this lost information. Thus we propose FINE and LORI which enables the modeling of long and out-of-range interactions. FINE serves as a generic module for windowed transformer-based models and exhibits promising preliminary results on BCV dataset. LORI is a versatile module for various deep learning methods. Experiments across three 3D medical image segmentation datasets, including CT-scans and ultrasound data, consistently demonstrate LORI's superiority over classical methods, underscoring its robustness and the importance of context modeling.

Keywords : Segmentation, Transformers, Computer Vision, Deep Learning, Medical Image.

Contents

Remerciements **3**

Résumé **7**

Abstract **11**

Liste des tableaux **16**

Liste des figures **22**

Résumé de la thèse **25**

1 Introduction **37**
1.1 Context 38
1.2 Motivations and challenges 41
1.3 Main trends in medical image segmentation 45
1.3.1 Medical image segmentation before deep learning 45
1.3.2 Convolutional neural networks (CNNs) 45
1.3.3 Transformers 47
1.4 Contributions and Outline 54
1.5 Related publications 57

2 U-Net Transformer: Self and Cross Attention for Medical Image Segmentation **59**
2.1 Introduction 61
2.2 Related Work 63
2.3 The U-Transformer Network 64
2.3.1 Self-attention 65
2.3.2 Cross-attention 65
2.4 Experiments 66
2.4.1 U-Transformer performances 67

2.4.2 U-Transformer analysis and properties . . . 69
2.5 Conclusion . . . 71

3 Full Contextual Attention for Multi-resolution Transformers in Semantic Segmentation 73
3.1 Introduction . . . 75
3.2 Related work . . . 76
3.3 The GLAM Method . . . 77
3.3.1 Global attention multi-resolution transformers . . . 78
3.4 Experiments . . . 82
3.4.1 Experimental Settings . . . 82
3.4.2 GLAM performance . . . 83
3.4.3 Additional results . . . 85
3.4.4 Model Analysis . . . 86
3.4.5 Visualizations. . . . 89
3.5 Conclusion . . . 91

4 LORI: Long and Out of Range Interaction transformer module for 3D medical image segmentation 95
4.1 Introduction . . . 97
4.2 Indirect attention modeling . . . 98
4.3 The Full resolutIoN mEmory transformer (FINE) . . . 99
4.4 Long-and-Out-of-Range Interaction method (LORI) . . . 101
4.5 Experiments . . . 106
4.5.1 Datasets . . . 106
4.5.2 Implementation details . . . 106
4.5.3 Backbones . . . 107
4.6 Results . . . 108
4.6.1 Preliminary results . . . 108
4.6.2 Results . . . 109
4.6.3 LORI Model analysis . . . 111
4.7 Conclusion . . . 115

5 Conclusion and Perspectives 117
5.1 Contributions . . . 118
5.2 On going work . . . 119
5.3 Perspectives for futures Works . . . 119

List of Appendices **136**

A Detailed Non Local Upsampling **137**

List of Tables

2.1 Results for each method in Dice similarity coefficient (DSC, %) on TCIA and IMO. Bold indicates best performances. 68
2.2 Extended results in Dice (%) on TCIA using nnU-Net [1] baseline and experimental setup. In this setup, U-Transformer has more layers, and the MHSA is applied also on more layers. The number of layers match nnU-Net architecture. 68
2.3 Results on IMO in Dice similarity coefficient (DSC, %) detailed per organ. 69
2.4 Ablation study on the positional encoding and multi-level on one fold of TCIA and IMO. 70
2.5 Hausdorff Distances (HD) for the different models 70

3.1 **GLAM Improvements on various multi-resolution transformers.** Performances are evaluated with respect to mIoU for ADE20k and Cityscapes and average DSC for BCV. 83
3.2 Comparison to state of the art methods on BCV. 84
3.3 Comparison to state of the art methods on ADE20K and Cityscapes. All experiments are made or reported are with single-scale inference. 85
3.4 **GLAM Improvements with Multi Scale inference on ADE20K.** Performances are evaluated with respect to mIoU for single scale inference (SS) and multiscales inference (MS). 86
3.5 **GLAM Improvements with Multi Scale inference on Cityscapes.** Performances are evaluated with respect to mIoU for single scale inference (SS) and multiscales inference (MS). 86
3.6 Detailed per-organ comparison on the multi-organ Synapse dataset (Dice Score in %). 86
3.7 Impact of the NLU and the GLAM transformer on a tiny Swin-UNet, 10 global tokens, on ADE20k. 87
3.8 Impact of G-MSA phase on GLAM transformer on different model, 10 global tokens, on ADE20k. GLAM-nogmsa is GLAM without the G-MSA phase. 87

3.9 Analysis of the relative mIoU increase with respect to extra learnable parameters and FLOPs compared to the standard Base and Large backbones. 88

3.10 Global token merging (tiny Swin-Unet, ADE20k). 89

4.1 Method comparison using the BCV dataset and the training / test split from [2]. Average Dice scores are shown (DSC in % - higher is better). The average and individual organ 95% Hausdorff distances are also shown (HD95 in mm - lower is better). * denotes results trained by us using the authors' public code. 108

4.2 Method comparison with SOTA transformer baselines (CoTr and nnFormer) using the BCV dataset and 5-fold cross validation. Results show mean and standard deviation of Dice (in %) for each organ and the average Dice over all organs (higher is better). 109

4.3 Ablation study of the impact of different tokens on BCV dataset. The metrics are Dice score (DSC in %) for all organs and in average, and the 95% Hausdorff distance (HD95 in mm). WT: Window tokens. VT: Volume tokens. 109

4.4 Comparison of the overall organs mean Dice score (in %) of LORI with state-of-the-art methods on three different datasets: WORD, BCV, and LIVUS. 110

4.5 Detailed results on WORD's dataset. The Dice score in % is given. The p-values between LORI and the methods we trained ourselves is also given. 110

4.6 Results on WORD's dataset showing 95% Hausdorff distance (HD96) and average symmetric surface distance (ASSD) metrics. 111

4.7 Study of the importance of long-range mechanism (GLAM), the combination of long and out-of-range mechanisms (LORI) and versatility of LORI module. This ablation shows dice score in % on LIVUS and WORD dataset. The number of parameter of each method is given showing a small increase compared to the gains. 112

List of Figures

1.1 Significant applications of AI research: Faces recognition [3] and video Action recognition [4] in computer vision, Audio speech to text [5] for audio analysis and text translation [6] for natural language processing. 39

1.2 Examples of traditional computer vision tasks: Classification with an image of a Border Collie from ImageNet [7], Object detection with the detection of multiple objects (cars, bicycle, truck) in a city [8], and Segmentation of a city [9] with multiple objects segmented (tree, sky, cars, pedestrian). 39

1.3 Example of medical image analysis modalities (CT: Computed Tomography; MRI: Magnetic Resonance Imaging; US: Ultrasound images). 41

1.4 3D Ultrasound image and segmentation of Liver and Vessels. The segmentation mask is done by an expert. These images come from a private dataset collected at IRCAD. 42

1.5 Challenges in US images interpretation: US images show multiple sources of data uncertainty. As shown in the image, there are often acoustic shadows which hide parts of the image. Compared to CT and MR, US usually has a higher signal-to-noise ratio, which reduces the capacity to find details as shown on the image's zoom. 42

1.6 Two examples of Ultrasound images with Liver and vessel segmentation by three experts. We see on the first row the disagreement to segment the IVC (in orange) with Expert 2 vs Expert 1 & 3. On the second row, there is a disagreement about segmenting the HV (in green) with Expert 2 vs Expert 1 & 3, and the liver (in blue) with Expert 3 vs. Expert 1 & 2. 43

1.7 **Long and out-of-range limitations.** This schema shows the out-of-range information loss effect due to random cropping, a commonly employed strategy for training segmentation models on large 3D medical images. On b) we see a zoom on the cropped patch (in red) from the original image (in green). The Receptive Field (in blue) in a CNN designates the area of the input image reached by a unit at the end of the network. When classifying the pixel in the center of the blue square, the area outside the receptive field is thus not taken into account. The information outside the patch in the original image is, by definition, not used during segmentation. Through this manuscript, we designate as long-range the information not encompassed by standard CNN backbones receptive field and as out-of-range areas outside the cropped patch. . 44

1.8 **AlexNet schema [10].** Simplified schema of AlexNet with the different succession convolution and pooling layers, followed by dense layers. 46

1.9 3D UNet architecture schema [11]. 47

1.10 **The transformer architecture.** This schema shows the detailed original transformer model architecture with the encoder and decoder parts and all its sub-modules: embedding of the input tokens, positional encoding, multi-head self-attention, normalization, skip connections, and feed-forward network. 49

1.11 **ViT model schema [12].** ViT splits an image into fixed-size patches, linearly embeds each of them, add position embeddings, and feeds the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, ViT use the standard approach of adding an extra learnable "class token" to the sequence. 50

1.12 Example of segmentation produced by SAM on US images [13]. We observe that SAM struggle to segment the breast tumour and the kidney. Even if SAM is trained on an extra large dataset and is design to be an universal segmentation model, it suffers from domain shift. 50

1.13 **Schema of CoTr [14].** A CNN-encoder, a DeTrans-encoder, and a decoder. Gray rectangles: CNN blocks. Yellow rectangles: 3D deformable Transformer layers. The CNN-encoder extracts multi-scale feature maps from an input image. The DeTrans-encoder processes the flattened multi-scale feature maps that embedded with the positional encoding in a sequence-to-sequence manner. The features with long-range dependency are generated by the DeTrans-encoder and fed to the decoder for segmentation. 52

1.14 Taxonomy of Efficient Transformer Architectures [15]. 53

1.15 **Linformer [16].** Left and bottom-right show architecture and example of Linformer multihead linear self-attention. Top right shows inference time vs. sequence length for various Linformer models. 54

1.16 **Comparison between windowed based transformer and ViT [17].** (a) Swin Transformer builds hierarchical feature maps by merging image patches (shown in gray) in deeper layers and has linear computation complexity to input image size due to computation of self-attention only within each local window (shown in red). It can thus serve as a general-purpose backbone for both image classification and dense recognition tasks. (b) In contrast, previous ViT-produced feature maps of a single low resolution and have quadratic computation complexity to input image size due to computation of self-attention globally. 55

2.1 Global context is crucial for complex organ segmentation but cannot be captured by vanilla U-Nets with a limited receptive field, *i.e.* blue cross region in a) with failed segmentation in c). The proposed U-Transformer network represents full image context by means of attention maps b), which leverage long-range interactions with other anatomical structures to properly segment the complex pancreas region in d). . 61

2.2 **The Effective Receptive Field as formulated in [18].** We put a gradient of one at the end of the encoder and propagate it to the input. The figures show high gradient values in white and zero gradients in black. We analyze the U-Net and nnU-Net architectures and observe that the final ERF is much smaller than the TRF. a) b) and c) have the same dimensions (512x512 pixels). 62
2.3 **The Attention U-Net as proposed in** [19]: The top image is the overall architecture with Attention Gates (AGs) at each skip connection. The bottom image is the attention gate mechanism with g being the gating signal (from the previous decoder block) and x the input signal (the skip connection). 64
2.4 **U-Transformer** augments U-Nets with transformers to model long-range contextual interactions. The Multi-Head Self-Attention (MHSA) module at the end of the U-Net encoder gives access to a receptive field containing the whole image (shown in purple), in contrast to the limited U-Net receptive field (shown in blue). Multi-Head Cross-Attention (MHCA) modules are dedicated to combine the semantic richness in high level feature maps with the high-resolution ones coming from the skip connections. 65
2.5 **MHSA module**: the input tensor is embedded into a matrix of queries Q, keys K and values V. The attention matrix A in purple is computed based on Q and K. (1) A line of A corresponds to the attention given to all the elements in K with respect to one element in Q. (2) A column of the value V corresponds to a feature map weighted by the attention in A. 66
2.6 **MHCA module**: the value of the attention function corresponds to the skip connection S weighted by the attention given to the high level feature map Y. This output is transformed into a filter Z and applied to the skip connection. 67
2.7 Segmentation results for U-Net [20], Attention U-Net [19] and U-Transformer on the multi-organ IMO dataset (first row) and on TCIA pancreas (second row). 69
2.8 Cross-attention maps for the yellow-crossed pixel (left image). 70
2.9 Evolution of the Dice Score on TCIA (fold 1) when the number of heads varies between 0 and 8 in MHSA. 71

3.1 When segmenting the high-resolution image in a) with state-of-the-art multi-resolution transformers, *e.g.* Swin [17], the attention in the highest-resolution feature maps is limited to a small spatial region, *i.e.* the blue square for the yellow-crossed pedestrian. Our method incorporates GLobal Attention in Multi-resolution transformers (GLAM). The GLAM attention map for the pedestrian in a) is depicted in b): it captures both fine-grained spatial information and long-range interactions, enabling successful segmentation, as shown in d). 76

3.2 The GLAM module for modeling full-range interaction in multi-resolution transformers. GLAM is included at each resolution level of any multi-resolution transformer architecture, *e.g.* Swin-UNet [17] or Swin-UperNet [17]. GLAM includes learnable global tokens, which are leveraged into a succession of two attention steps. We show that this design can indirectly represent long-range interactions between all image regions at all scales, and also external information useful for segmentation while retaining efficiency. We also introduce a non-local upsampling scheme (NLU) to extend the global context modeling in full transformer U-shape architectures such as [21, 2]. 78

3.3 **GLAM-Transformer:** as in multi-resolution approaches, each input feature map is divided into N_w non overlapping windows (blue). The core idea in GLAM is to design learnable global tokens (in red). The visual tokens from each window are concatenated with the global tokens and processed through a local window transformer (W-MSA). Every W-MSA is followed by a global transformer (G-MSA), where global tokens between different windows interact with each other, which brings a global representation to each window. These two steps give the GLAM-Transformer block; Multiple blocks are chained at every hierarchy level in typical multi-resolution transformer backbones. We show that global tokens learned from GLAM-Transformer indirectly model global interactions between all visual tokens in all widows. The global tokens are also able to represent extra learnable knowledge beyond the patch interactions in a single image. 79

3.4 Impact of the number of global tokens on performance (mIoU) using ADE20k. . . . 87

3.5 **Averaged GLAM attention map in 3D.** The information inside the blue window is ambiguous. To segment the voxel at the red cross, the model leverages long-range dependencies including neighbor organs. The pancreas is in green, the aorta in red, and the stomach in blue. 88

3.6 Qualitative visualisations of GLAM. We show the ability of GLAM to model full contextual information in high-resolution feature maps on ADE20K (first row), and the ability of GLAM-nn-Former to accurately segment the stomach (in pink). 90

3.7 Global attention of GLAM compared to vanilla Swin on ADE20K. 90

3.8 **Qualitative results of GLAM incorporated to Swin-upernet on Cityscapes** For two test images, we show from top-left to bottom-right : the image, the global attention map with respect to the blue window, the ground truth and the predicted segmentation. 91

3.9 **Qualitative results of GLAM incorporated to Swin-UNet on ADE20K** For two test images, we show from top-left to bottom-right : the image, the global attention map with respect to the blue window, the ground truth and the predicted segmentation. . . 92

3.10 **Qualitative results of GLAM Incorporated to nnFormer on BCV.** The observed organs are the liver (pink), the stomach (purple), the aorta (cyan) and the spleen (blue). 93

4.1 Visualization of LORI's Indirect Attention on the WORD Dataset. The red cross indicates the focal pixel, with the attention map highlighting its corresponding attention. Unlike the limitation imposed by the window size (blue square), LORI efficiently computes long-range attention across the entire input cropped image (green square). Furthermore, LORI effectively captures out-of-range attention from the complete volume, as demonstrated through visualizations on axial, sagittal, and coronal planes. These visualizations demonstrate LORI's capacity to harness information from diverse spatial dimensions, underscoring its potential in significantly enhancing 3D medical image segmentation. 98

4.2 To segment the cropped patch in blue and model global context, two levels of memory tokens are introduced: window (red) and volume (green) tokens. First, the blue crop is divided into windows over which Multi-head Self-Attention (MSA) is performed in parallel. For each window, the sequence of visual tokens (blue) is augmented with a specific window token. Second, the local information embedded into each window token is shared between all window tokens and volume tokens intersecting with the crop (light green). Finally, high-level information is shared between all volume tokens). 100

4.3 Overview of the LORI module: (a) shows the whole set of global tokens of LORI (purple tokens), each one associated with a region of the image. When a cropped patch is selected into the full-size image (red square), the associated global tokens are selected (red tokens). The set of visual tokens (yellow tokens) is also selected and reorganized into a subset of windows. (b) shows the first attention step of LORI where the selected global tokens are duplicated to form the global feature map and to be associated with each window. A window transformer (W-MSA) is applied to this feature map to share information between the global feature map and the visual tokens. (c) shows the second attention step of LORI done by a classic transformer (G-MSA) on the features composed of the global feature map which captured information from each window and the set of all global tokens. This step lets the global feature map share information between all windows and incorporates information from the full-size image. 102

4.4 Visualisation of segmentation masks produced by the different methods on the test set of WORD dataset. 104

4.5 This ablation study shows the dice score (in %) on the WORD dataset of LORI with different architecture parameters. In (a) we experimented LORI with different number of global tokens associated with each region of the full-size image. In (b) we experimented with different number of LORI modules in a layer. In (c) we experimented different position of LORI in the based model: 1 correspond to only the last layer; 2 correspond to the two last layers; 3 correspond to the three last layers; 4 correspond to the four last layers. 113
4.6 Visualisation of segmentation masks produced by the different methods on the test set of WORD dataset. 114
4.7 Visualisation of segmentation masks produced by the different methods on the test set of LIVUS dataset. 114

5.1 Qualitative results showing example CT and US segmentations using IRCAD's new dataset (patient 01 and 17). The top two rows shows coronal CT slices, with ground-truth segmentations overlaid in green, and estimated segmentations in blue. The bottom two rows shows longitudinal US slices, with ground truth segmentations overlaid in red, and estimated segmentations in blue. The two rows last (Annotator 1 and 2) show segmentations from each annotator, and the remaining rows show the best training version on average between single or double target(s) of each automatic segmentation from 5 DNN-based methods. 120

A.1 **Non-Local Upsampling** The upsampling is processed window by window and is conceived as a super-resolution module where the low resolution feature map in the decoder (red) are re-embedded based on the high resolution ones coming from the encoder (blue). The patches are downsampled by a factor 2 before each hierarchy in the models. A given region from the decoder corresponds then to four neighbouring windows in the feature map coming from the skip connection. 138

Résumé de la thèse

Segmentation sémantique d'images médicales 3D par deep learning

L'intelligence artificielle (IA), et en particulier l'apprentissage profond (Deep Learning, DL), a connu d'importantes avancées, impactant divers domaines tels que l'analyse d'images et de vidéos, la reconnaissance audio ou la traduction de texte. Ces progrès sont dus à deux facteurs principaux : le développement de vastes bases de données (par exemple, ImageNet pour la classification d'images, ADE20K pour la segmentation d'images, et COCO pour la détection d'objets) et l'augmentation du nombre de paramètres des modèles grâce aux progrès du calcul sur unité de traitement graphique (GPU). De plus, des techniques novatrices comme les modèles Transformers ont permis d'exploiter efficacement ces grandes bases de données. Ces avancées ont conduit à la création de modèles dits "Fondation" avec un nombre considérable de paramètres. Cela a mené au développement d'outils logiciels accessibles au grand public tels que Chat GPT et DALL-E, qui assistent dans diverses tâches comme la traduction, la programmation et la création d'images artistiques à partir de textes. Parallèlement, l'IA progresse dans des domaines spécifiques comme la climatologie, l'astronomie et la médecine, et est également utilisée dans le développement de jeux vidéo et de systèmes de conduite autonome.

La vision par ordinateur (Computer Vision, CV), un sous-domaine de l'intelligence artificielle, se concentre sur le traitement et l'analyse d'informations visuelles provenant de diverses sources telles que les caméras, les images et les vidéos. Elle comprend des tâches telles que la classification d'images, la détection d'objets et la segmentation d'images. La classification d'images consiste à attribuer une étiquette ou une classe spécifique à une image, en identifiant et catégorisant les principaux objets ou caractéristiques présents. La détection d'objets vise à déterminer la position précise d'un ou plusieurs objets dans une image, en identifiant différents objets et leurs coordonnées spatiales. La segmentation d'images, une autre tâche cruciale, implique d'assigner une classe spécifique à chaque pixel de l'image, permettant ainsi de diviser l'image en régions ou segments distincts, chacun associé à une catégorie particulière. Cette segmentation est particulièrement importante dans l'analyse d'images médicales.

Cette thèse se consacre à l'amélioration de l'analyse d'images médicales grâce à des innovations

en vision par ordinateur. Elle est réalisée en collaboration avec le Conservatoire Nationale des Arts et Métiers (CNAM) à Paris et l'Institut de Recherche sur les Cancers de l'Appareil Digestif (IRCAD) à Strasbourg. L'IRCAD est reconnu mondialement pour son excellence en recherche médicale. Cette thèse fait partie du projet Disrumpere[1] : Démocratisation du diagnostic automatique, du dépistage, de la biométrie et de la chirurgie percutanée augmentée assistée par l'intelligence artificielle. L'objectif principal de Disrumpere est d'augmenter l'accès aux soins de santé en utilisant l'intelligence artificielle pour améliorer les capacités des sondes à ultrasons portables et abordables, rendant leur utilisation accessible aux non-experts. Conduit par IRCAD France et IRCAD Afrique, le projet rassemble des équipes d'ingénieurs et de chercheurs déterminés à développer un outil pouvant impacter significativement la santé en Afrique et répondre aux défis des déserts médicaux en France. Le projet comporte plusieurs composants clés. Le premier implique le développement d'algorithmes performants pour le diagnostic et le suivi, facilitant la détection efficace et précise de pathologies courantes. Le deuxième se concentre sur la démocratisation de la chirurgie percutanée augmentée par la robotique, en utilisant des guidages par ultrasons pour effectuer des biopsies ou détruire de petites tumeurs cancéreuses. Disrumpere vise à rendre cette technique chirurgicale plus accessible, élargissant la portée des procédures peu invasives et améliorant les résultats pour les patients.

L'application de l'IA dans l'imagerie médicale est reconnue comme un domaine à fort potentiel bénéfique pour la société. L'IA peut assister les professionnels de santé dans diverses tâches, allant du diagnostic à la planification et à la guidance chirurgicale. Dans le domaine de l'analyse d'images médicales, une gamme variée de modalités d'imagerie médicale existe, chacune offrant ses propres avantages. Ces modalités incluent des images telles que des images rétiniennes ou cellulaires microscopiques, l'Imagerie par Résonance Magnétique (IRM), la Tomodensitométrie (CT) et l'Échographie (US). L'IRM offre une précision exceptionnelle pour capturer des informations détaillées sur des tissus mous tels que le cerveau, le cœur ou les tumeurs, de manière non invasive. La CT, quant à elle, utilise des rayons X pour obtenir des scans 2D ou 3D précis du corps entier, y compris les tissus et les os. L'échographie, une modalité non invasive et économique, fournit des images médicales en temps réel et sans les risques des radiations ionisantes. Avec la disponibilité de ces diverses modalités d'imagerie, de nombreuses tâches peuvent être accomplies en utilisant l'IA, y compris la détection de tumeurs ou de lésions, la segmentation d'images, la reconstruction d'images, et la segmentation d'organes pour aider au diagnostic ou à la planification de la chirurgie [22, 23, 1, 24].

La segmentation des images échographiques est une tâche complexe confrontée à des défis importants, principalement en raison de deux facteurs clés : la disponibilité limitée de données annotées et la qualité intrinsèque associée à la modalité des ultrasons (US). Comme indiqué, les images US sont intrinsèquement bruitées, avec un bruit de speckle particulièrement problématique dans le processus

[1] https://www.ircad.fr/fr/newsletter-de-lircad-decembre-2022/

de segmentation. De plus, la résolution des images US varie au sein du volume, entraînant des distorsions dans l'image. Certains tissus ont la capacité d'absorber complètement l'onde ultrasonore, ce qui entraîne des occlusions et des ombres dans les images. Ces défis limitent l'efficacité des modèles d'apprentissage profond conventionnels pour la segmentation d'images US. Ces artefacts posent des difficultés même pour les cliniciens experts, comme en témoignent les écarts dans les résultats de segmentation entre différents experts. En conséquence, les modèles d'apprentissage profond traditionnels rencontrent des limitations lorsqu'ils sont appliqués aux modalités US, avec des erreurs sur les bords et des confusions sur des tâches difficiles comme la segmentation de plusieurs types de vaisseaux. Par conséquent, il existe un intérêt croissant pour des stratégies qui exploitent les informations contextuelles pour surmonter ces défis dans la communauté de recherche. En incorporant un contexte supplémentaire, la capacité du modèle à classer avec précision un pixel bruyant ou occulté peut être améliorée en considérant les dépendances à longue portée. Dans les régions plus bruyantes ou ombragées, le modèle doit étendre son analyse au-delà du voisinage immédiat afin de discerner les structures prédominantes dans ladite région.

L'utilisation d'images 3D en imagerie médicale a apporté des avantages significatifs, permettant une modélisation contextuelle plus complète. L'utilisation de volumes plutôt que de tranches 2D est une approche supérieure car elle fournit plus d'informations au modèle lors de la segmentation de voxels. Avec la capacité d'examiner des structures sous plusieurs directions, le modèle peut mieux exploiter le contexte dans la segmentation des images US, CT ou IRM. Bien qu'il existe plusieurs architectures qui exploitent directement les volumes, elles nécessitent souvent une complexité spatiale élevée, entraînant une consommation significative de mémoire GPU. L'utilisation d'une méthode de pointe telle que 3D-UNet [11] rend impossible le traitement de l'ensemble du volume de l'image. Les images 3D ont souvent de grandes dimensions, mais même l'utilisation de la première couche de cette image 3D consommerait environ 24 Go de mémoire GPU. De plus, le modèle entier nécessiterait plus de 60 Go de mémoire GPU pour le traitement d'un seul volume lors de l'inférence, et plus de 120 Go lors de la formation. Cette contrainte matérielle, due au coût élevé des GPU, peut avoir un impact négatif sur la capacité et les performances des modèles de DL. Pour résoudre ce problème, les chercheurs ont adopté une stratégie commune consistant à entraîner des modèles DL sur une région plus petite du volume original [1, 25], permettant au modèle de traiter des portions plus petites du volume pendant la formation et l'inférence. Cette méthode permet au modèle de segmenter le volume complet en utilisant une stratégie de fenêtre glissante lors de l'inférence. Cependant, la stratégie commune est de travailler avec des patchs de volume découpés. Une configuration typique serait, pour un ensemble de données médicales, de montrer une taille moyenne de volume de $512 \times 512 \times 256$ voxels et de s'entraîner sur des cultures aléatoires de dimension $128 \times 128 \times 64$, ce qui ne représente que 1.6 % du volume original. Cela implique une perte de l'information et du contexte total disponibles

pour le modèle DL à exploiter. À cause de cette stratégie, les méthodes perdent ce que nous appelons des informations hors de portée, c'est-à-dire des informations en dehors de l'image découpée. De plus, même dans les patchs découpés, la plupart des méthodes ne peuvent pas modéliser spécifiquement les interactions à longue portée dans les caractéristiques de haute résolution.

Cette thèse a pour objectif d'étudier de nouvelles méthodologies pour améliorer la précision de la segmentation des images médicales 3D. **L'accent est mis sur l'exploration d'approches qui utilisent efficacement le contexte global dans les images 3D. Cela signifie modéliser les interactions à l'intérieur et à l'extérieur des patchs découpés, tout en maintenant une haute résolution spatiale pour exploiter davantage d'informations et, par conséquent, obtenir une segmentation plus précise.**

La première partie présente le U-Transformer, une innovation qui intègre le mécanisme d'attention des Transformers dans une architecture UNet pour la segmentation d'images médicales 2D. Cette approche résout le problème du champ de réception limité rencontré avec les architectures UNet classiques. Le U-Transformer, pionnier dans l'application des Transformers à la segmentation d'images médicales, se distingue par sa capacité à modéliser une attention globale dans le goulot d'étranglement de l'encodeur, contrairement aux modèles d'attention standard (*exemple* Attention-UNet [19]) qui ne renforcent pas le contexte global. C'est aussi l'une des premières utilisations des Transformers en vision par ordinateur après la publication de ViT [12]. Le U-Transformer surmonte les limitations des U-Nets traditionnels, notamment leur incapacité à modéliser des interactions contextuelles à longue portée et des dépendances spatiales essentielles pour une segmentation précise dans des contextes difficiles. Il intègre des mécanismes d'attention à deux niveaux : un module d'auto-attention qui exploite les interactions globales entre les caractéristiques de l'encodeur, et une attention croisée dans les "skip connexions" qui améliore la récupération spatiale fine dans le décodeur UNet en filtrant les caractéristiques non sémantiques. Les expériences montrent une nette amélioration des performances du U-Transformer par rapport au U-Net classique et au Attention U-Net local. L'article souligne également l'importance de combiner auto- et attention croisée, ainsi que les capacités d'interprétabilité offertes par le U-Transformer.

La deuxième partie de la thèse aborde la difficulté que rencontre le U-Transformer à gérer la haute dimensionnalité des images médicales 3D, due à la complexité quadratique de l'auto-attention et de l'attention croisée. Cette complexité limite l'efficacité du modèle à traiter les interactions à longue portée avec des caractéristiques de haute résolution. Pour résoudre ce problème, notamment dans le traitement d'images haute résolution et d'images 3D, la deuxième phase de cette thèse introduit GLAM (GLobal Attention Multi-resolution transformers). GLAM, un module générique pouvant être intégré dans la plupart des architectures de Transformers existantes, se distingue par l'inclusion de tokens globaux apprenables. Ces tokens, contrairement aux méthodes précédentes, peuvent modéliser

des interactions entre toutes les régions de l'image et extraire des représentations puissantes durant l'entraînement. Inspirés par le token de classe [26, 12], ils jouent un rôle clé dans la transmission d'informations à travers différentes régions de l'image à chaque étape. L'intégration de GLAM dans le modèle permet des interactions étendues sur de longues distances, absentes auparavant. Des expériences approfondies révèlent que GLAM surpasse nettement les performances des modèles état de l'art sur des images 2D de grandes dimensions. De plus, GLAM montre également de bonnes performances sur des images médicales 3D.

La troisième partie explore une nouvelle approche pour modéliser des informations contextuelles complètes, y compris des interactions hors de portée, lors de l'entraînement de modèles à partir de patches 3D locaux. Cette méthode vise à surmonter la limitation de ne pouvoir exploiter des informations au-delà des frontières du volume d'entrée. S'appuyant sur le concept de GLAM, qui utilise des tokens globaux pour modéliser indirectement des informations, cette approche est étendue pour inclure la modélisation d'interactions hors de portée, même dans des cartes de caractéristiques à haute résolution. Notre contribution est, à notre connaissance, la première à intégrer des informations au-delà du volume d'entrée coupé dans le contexte de la segmentation d'images médicales 3D. Nous introduisons une méthode visant à résoudre les défis mentionnés en facilitant l'intégration de dépendances à longue portée et hors de portée dans les modèles de segmentation médicale. Cette méthode intègre des tokens globaux et utilise des mécanismes d'auto-attention pour créer des interactions à longue portée et hors de portée. Deux variantes de cette méthode sont proposées : FINE (Full resolutIoN mEmory transformer), une architecture entièrement basée sur les transformateurs servant de preuve de concept préliminaire, et LORI, un module générique pouvant être intégré sans problème dans des modèles existants tels que nnUNet. Des expériences préliminaires sur BCV avec FINE démontrent sa pertinence, et des évaluations expérimentales approfondies avec LORI sur trois ensembles de données distincts: deux ensembles de données de segmentation multi-organes CT 3D et un ensemble de données d'images échographiques 3D pour la segmentation du foie et des vaisseaux. Les résultats obtenus montrent une amélioration substantielle des performances de segmentation avec LORI. Notamment, LORI a montré des performances supérieures sur plusieurs ensembles de données d'images 3D haute résolution multi-classes, indépendamment des différentes modalités impliquées.

U-Net Transformer: Self and Cross Attention for Medical Image Segmentation

Jusqu'à récemment, les méthodes de pointe en matière de segmentation d'images s'appuyaient sur des Réseaux Convolutifs (Fully Convolutional Networks, FCNs), tels que U-Net et ses variantes [20, 11, 27, 28]. Les architectures U-Net, basées sur un schéma encodeur-décodeur, extraient des représentations sémantiques de haut niveau via une cascade de couches convolutionnelles. Le décodeur, quant à lui, utilise des "skip connexions" pour réutiliser les cartes de caractéristiques à haute résolution de l'encodeur, dans le but de récupérer les informations spatiales perdues dans les

représentations de haut niveau. Malgré leurs performances exceptionnelles, les FCNs présentent des limites conceptuelles dans les tâches de segmentation complexes, notamment lorsqu'il s'agit de gérer des ambiguïtés visuelles locales et un faible contraste entre les organes. Les structures présentant des dépendances spatiales à longue portée dans des régions à faible contraste peuvent entraîner une mauvaise classification. De plus, la classification de petits organes ou d'organes à variabilité de forme significative, tels que le pancréas, représente un défi supplémentaire.

Dans cette partie, nous présentons le réseau U-Transformer, qui tire parti des capacités des transformers pour modéliser les interactions à longue portée et les relations spatiales entre les structures anatomiques. Le U-Transformer conserve le biais inductif de la convolution grâce à son architecture en forme de U, mais introduit des mécanismes d'attention à deux niveaux, ce qui aide à modéliser le contexte global et à interpréter les décisions du modèle. Premièrement, un module d'auto-attention exploite les interactions globales entre les caractéristiques sémantiques à la fin de l'encodeur pour modéliser explicitement l'information contextuelle complète. Deuxièmement, nous introduisons une attention croisée dans les "skip connexions" pour filtrer les caractéristiques non sémantiques, permettant une récupération spatiale fine dans le décodeur U-Net.

Comme mentionné précédemment, les architectures en forme de U de type encodeur-décodeur manquent d'informations contextuelles globales pour gérer des tâches complexes de segmentation d'images médicales. Pour remédier à cela, nous introduisons le réseau U-Transformer, qui enrichit les U-Nets avec des modules d'attention basés sur des transformers à têtes multiples. Le U-Transformer modélise les interactions contextuelles à longue portée et les dépendances spatiales en utilisant deux types de modules d'attention : l'Auto-Attention à Têtes Multiples (Multi-Head Self-Attention, MHSA) et l'Attention Croisée à Têtes Multiples (Multi-Head Cross-Attention, MHCA). Ces deux modules sont conçus pour exprimer une nouvelle représentation de l'entrée basée, dans le premier cas, sur son auto-attention ou, dans le second cas, sur l'attention portée aux caractéristiques de niveau supérieur. Ces modules permettent au U-Transformer de surmonter les limitations des architectures U-Net traditionnelles en fournissant un cadre amélioré pour la compréhension et la segmentation des images médicales complexes.

Le module MHSA est conçu pour extraire des informations structurales à longue portée des images. Pour ce faire, il est composé de fonctions d'auto-attention à têtes multiples, comme décrit dans les travaux de Vaswani et al. [29], et est positionné dans le goulot d'étranglement du U-Net. L'objectif principal du MHSA est de connecter chaque élément de la carte de caractéristiques sémantiquement riches avec tous les autres, offrant ainsi un champ de réception qui englobe toute l'image d'entrée. La décision concernant un pixel spécifique peut donc être influencée par n'importe quel pixel de l'entrée.

L'attention peut également être utilisée pour augmenter l'efficacité du décodeur U-Net, en particulier pour améliorer les cartes de caractéristiques de niveau inférieur transmises via les "skip

connexions". Bien qu'elles conservent des informations à haute résolution, elles manquent de la richesse sémantique présente plus profondément dans le réseau. L'idée derrière le module d'Attention Croisée à Têtes Multiples (MHCA) est de désactiver les zones bruitées ou non pertinentes des caractéristiques de la "skip connexions" et de mettre en évidence les régions présentant un intérêt significatif pour la tâche.

L'évaluation de U-Transformer a été réalisée pour la segmentation d'organes abdominaux sur deux ensembles de données : le dataset public du pancréas de The Cancer Imaging Archive (TCIA) [30] et un ensemble de données interne multi-organes (IMO). La segmentation précise du pancréas est particulièrement difficile en raison de sa petite taille, de sa forme complexe et variable, et du faible contraste avec les structures avoisinantes. De plus, le contexte multi-organes permet d'évaluer comment U-Transformer peut tirer parti de l'attention provenant des annotations de plusieurs organes. Pour une comparaison équitable, U-Transformer a été comparé au modèle de base U-Net [20] et à l'Attention U-Net [19] qui possèdent le même fond convolutif. Les performances ont également été évaluées en utilisant uniquement MHSA et uniquement l'attention croisée MHCA. U-Net compte environ 30 millions de paramètres, et l'augmentation de paramètres due à U-Transformer est limitée (environ 5 millions pour MHSA, et environ 2,5 millions pour chaque bloc MHCA).

Le U-Transformer a surpassé U-Net de 2.4 points sur le dataset TCIA et de 1.3 points pour IMO, et a également dépassé l'Attention U-Net de 1.7 points pour TCIA et de 1.6 points pour IMO. Des "paired t-tests" montrent que l'amélioration est significative avec des "p-values" inférieures à 3% pour chaque expérience. Des expériences supplémentaires ont été menées en utilisant nnU-Net [1] comme modèle de base et pipeline d'entraînement. nnU-Net, une version plus robuste et plus profonde de U-Net, est bien optimisé et entraîné dans un pipeline d'entraînement sur mesure pour la segmentation d'images médicales. Il obtient les meilleurs résultats sur plusieurs tâches. Notre approche a été évaluée en utilisant le code Github des auteurs de nnU-Net sur TCIA avec un pliage en 3 et en suivant la configuration expérimentale de [1]. Nos résultats montrent un gain d'environ 1 point (84.08 contre 83.09 en indice de Dice), ce qui est une amélioration importante étant donné la forte baseline, et statistiquement significatif avec un "paired t-test" (p=0.023). Cela souligne que nos modules MHSA/MHCA améliorent les performances par rapport aux modèles convolutionnels de pointe.

Full Contextual Attention for Multi-resolution Transformers in Semantic Segmentation

Le principal attrait des transformers réside dans leur capacité à saisir des interactions à longue portée, un élément crucial pour la segmentation sémantique. Cependant, cette stratégie n'est pas facilement extensible aux images haute résolution impliquant un grand nombre de patches, en raison de la complexité quadratique du module d'attention des transformers. Par exemple, le U-Transformer présenté applique uniquement l'auto-attention dans le goulot d'étranglement, et l'attention croisée

utilise une opération de sous-échantillonnage qui dégrade l'information spatiale. Une stratégie simple et efficace pour surmonter cette limitation est de s'appuyer sur des approches multi-résolutions, où l'attention dans les cartes de caractéristiques haute résolution est calculée sur des sous-fenêtres. Plusieurs tentatives récentes ont été faites dans cette direction. Cependant, elles limitent les interactions des caractéristiques haute résolution à l'intérieur de chaque fenêtre. Nous introduisons une approche pour la segmentation sémantique qui incorpore une attention globale dans les transformers multi-résolutions (GLAM). Le module GLAM permet de modéliser des interactions longue portée à toutes les échelles d'un transformer multi-résolution. L'incorporation de GLAM dans l'architecture Swin [17] permet de capturer conjointement des informations spatiales détaillées dans des cartes de caractéristiques haute résolution et un contexte global, deux éléments cruciaux pour une segmentation appropriée dans des scènes complexes.

L'idée principale de GLAM est de fournir un moyen de représenter des interactions complètes à toutes les résolutions de cartes de caractéristiques, ce qui est impossible dans les modèles classiques, en particulier dans les cartes de caractéristiques à haute résolution, en raison de la complexité quadratique de l'attention des transformers. Il est important de noter que GLAM peut être intégré dans diverses architectures multi-résolutions, par exemple Swin [21] ou en segmentation 3D, par exemple nnFormer [2]. L'idée centrale de GLAM est de concevoir des tokens globaux, qui sont utilisés dans une succession de deux étapes d'attention : d'abord, entre les tokens visuels dans chaque fenêtre indépendamment, et ensuite, entre les tokens globaux parmi différentes fenêtres. Nous montrons que cette conception permet de représenter des interactions complètes entre toutes les régions de l'image à toutes les échelles, et également des informations externes utiles pour la segmentation tout en conservant l'efficacité. Nous introduisons également un module de sur-echantillonage non local (NLU) pour étendre la modélisation du contexte complet dans les architectures en forme de U et pour fournir une interpolation efficace de cartes de caractéristiques sémantiques riches dans un décodeur associé. L'idée de base derrière GLAM est d'associer des tokens globaux à chaque fenêtre, chargés de capturer l'information locale et de la transmettre à d'autres régions de l'image en calculant une MHSA entre tous les tokens globaux. Ainsi, lorsque l'information est traitée à l'échelle de la fenêtre, l'encodage des tokens visuels intègre des informations utiles à longue portée. La communication entre les fenêtres à un niveau hiérarchique donné dans le transformer GLAM est obtenue grâce à l'interaction des tokens globaux. À chaque bloc l du transformer GLAM, il y a deux étapes : *i)* les tokens visuels captent leurs statistiques locales à travers un transformer à fenêtre locale (W-MSA), et *ii)* les tokens globaux sont ré-encodés par un transformer global (G-MSA), où les tokens globaux de différentes fenêtres interagissent entre eux. Formellement, le l^{me} bloc du transformer GLAM prend en entrée z^{l-1} et produit en sortie z^{l} par la succession d'une étape W-MSA et d'une étape G-MSA.

GLAM a été évaluée sur trois ensembles de données distincts pour la segmentation sémantique

: ADE20K [31], Cityscapes [32] et BCV [33]. ADE20K est un ensemble de données de parsing de scènes composé de 20 210 images réparties en 150 classes d'objets. Cityscapes contient des scènes de conduite et comprend 5 000 images annotées avec 19 classes différentes. BCV est un ensemble de données pour la segmentation d'organes abdominaux, incluant 30 scans CT qui sont des volumes 3D annotés avec 8 organes abdominaux différents. En raison des performances supérieures de Swin, GLAM a été intégré à cette base pour la segmentation de jeux de données 2D, résultant en deux modèles : GLAM-Swin-UperNet et GLAM-Swin-UNet. Le premier est un modèle hybride combinant une base de transformers et une tête de CNN, tandis que le second est un modèle transformers complet avec un décodeur symétrique à l'encodeur. Pour les images 3D, GLAM a été intégré dans nnFormer, conçu de manière similaire à Swin-UNet pour la segmentation d'images médicales 3D. Les performances des modèles Swin et GLAM montrent des gains significatifs et constants par rapport à leurs versions originales, que ce soit sur des modèles plus petits ou plus grands, avec environ +1.5 point sur ADE20K avec Swin-UNet, et +1.2 point sur BCV avec le modèle nn-Former. GLAM-nnFormer surpasse significativement toutes les autres méthodes médicales existantes avec au moins 1.2 points de Dice en plus en moyenne. À notre connaissance, GLAM-nnFormer dépasse l'état de l'art sur l'ensemble de données BCV. De plus, GLAM-Swin-UNet atteint 49.10% de mIoU sur ADE20K, surpassant son homologue Swin vanilla d'au moins 1.10 points de mIoU. GLAM-Swin-UperNet obtient 81.47% de mIoU sur Cityscapes, ce qui est 1.58 points de mieux que son homologue Swin-Upernet.

LORI: Long and Out of Range Interaction transformer module for 3D medical image segmentation

Dans le domaine actuel de la segmentation basée sur le DL, la plupart des méthodes existantes [34, 14, 35, 36] ne peuvent pas traiter des images médicales 3D complètes et sont limitées à traiter des sous-régions de l'image d'entrée, c'est-à-dire des patches extraits plus petis. Contrairement aux tranches 2D, ces patches préservent la nature 3D de l'entrée, tout en conservant la résolution originale du volume et en maintenant tous les détails fins. Cependant, cette approche présente des inconvénients. En effet, les patches sont traités indépendamment, ce qui entraîne une perte dramatique de contexte : les informations en dehors du patch, c'est-à-dire les informations hors de portée, ne peuvent pas être utilisées dans la prédiction et sont perdues. Par conséquent, lors de scénarios de segmentation difficiles, par exemple des organes complexes ou des données bruyantes, les modèles ont souvent du mal à produire une segmentation précise. L'objectif principal de cette partie est de résoudre ce problème en utilisant une nouvelle méthode pour modéliser les interactions hors de portée. Dans cette partie, nous généralisons le concept de tokens globaux comme pivot pour diffuser des informations multi-échelles dans l'attention. Comparé à l'auto-attention standard appliquée sur un volume 3D brut, cette approche offre un moyen de modéliser le contexte global tout en maintenant l'utilisation de la mémoire et le coût de calcul sous contrôle. Ainsi, nous présentons le transformer Long and

Out-of-Range Interaction (LORI).

Notre but est de modéliser des interactions à grande échelle et à haute résolution en généralisant le concept d'indirection de l'information via des tokens globaux, nécessitant l'identification de trois niveaux d'information. Au niveau de la fenêtre, l'objectif est de conserver des détails fins en calculant une auto-attention quadratique complète. Le deuxième niveau concerne le patch extrait, où l'on vise à propager l'information des fenêtres locales entre elles. Enfin, au niveau global, l'information provient du volume global, décrivant les structures de haut niveau dans l'image. Idéalement, l'information devrait circuler du niveau global au niveau fenêtre. Pour cela, chaque niveau est subdivisé en sous-régions, chacune associée à des tokens globaux dédiés, permettant une gestion flexible et efficace de l'information à différentes échelles.

LORI utilise des tokens globaux ancrés dans des régions du volume d'entrée. Lors d'une passe avant, les tokens globaux liés aux régions qui se chevauchent avec le patch recadré sont injectés dans le modèle. À travers LORI, ces tokens globaux font circuler l'information dans le patch et infusent des informations hors de portée. Ces tokens globaux agissent comme des représentations locales de parties anatomiques spécifiques. Le fait que toutes ces parties ne soient pas disponibles lors de la segmentation d'un patch donné représente un défi pour apprendre ces représentations. Plutôt que d'apprendre deux ensembles séparés de tokens globaux dédiés à extraire des représentations utiles de la structure sous-jacente et à propager des informations à haute résolution entre les fenêtres, LORI utilise un seul niveau de tokens globaux ancrés à chaque sous-région du volume. Ces tokens globaux sont appris de manière asynchrone en mettant à jour uniquement ceux associés aux régions qui se chevauchent avec les patches d'entraînement. L'injection de tokens globaux ancrés dans le module GLAM confère au modèle la capacité d'utiliser les représentations apprises des régions entourant le patch. Cette information permet au modèle d'aligner les tokens visuels de l'entrée avec les représentations de haut niveau apprises, améliorant ainsi leur pertinence pendant l'entraînement. Allant plus loin, en enchaînant de multiples opérations W-MSA et G-MSA, LORI utilise non seulement les informations environnantes, mais aussi les représentations de la structure sous-jacente dans son ensemble, permettant au modèle de saisir indirectement les interactions au-delà du patch observé. La séquence complète des opérations de LORI est donnée par une modification du GLAM introduit précédemment. LORI a la capacité de reproduire l'échange d'informations entre les fenêtres effectué par la modélisation des interactions à longue portée. De plus, via G-MSA, l'attention entre les tokens globaux sélectionnés et tous les tokens globaux est évaluée, permettant ainsi un partage indirect d'informations de l'ensemble du volume. Par la suite, dans le bloc suivant, les informations hors de portée recueillies par les global tokens sélectionnés sont transmises aux tokens visuels de la patch extrait, partageant ainsi les informations capturées.

Pour démontrer l'efficacité de la méthode proposée, des expériences ont été menées sur trois

ensembles de données distincts : WORD [37] un ensemble de donnée de segmentation multi-organes CT scans 3D, la base de segmentation multi-organes CT Scans 3D Synapse (BCV) [38], et un ensemble de données privé de segmentation du foie et de ces vaisseaux en 3D nommé LIVUS. Pour chaque ensemble de données, les résultats ainsi que leurs écarts-types correspondants ont été rapportés, ces derniers ayant été calculés sur l'ensemble des différents patients présents dans les ensembles de données. Dans cette étude, l'efficacité de l'approche proposée a été évaluée en la comparant à diverses architectures d'état de l'art (SOTA) : CNN avec nnUNet [34] et DeepLabV3 (2D) [39], Transformer avec 3D Swin-UNet [35] et FINE [40], ainsi que des hybrides avec CoTr [14] et UNETR [36]. À l'exception de DeepLabV3 et UNETR, tous les modèles ont été entraînés dans la même configuration, pour des fins de comparaison. LORI, étant un module polyvalent, peut être intégré sans problème dans divers modèles de segmentation. Dans cette étude, nnUNet a été choisi comme base pour LORI en raison de ses performances SOTA sur plusieurs ensembles de données. De plus, l'architecture basée sur la convolution de nnUNet présente une limitation en termes de petits champs de réceptif au niveau des cartes de caractéristiques à haute résolution. Cependant, cette limitation est efficacement abordée par LORI, qui étend la capacité du modèle à capturer des interactions à longue portée. Pour intégrer LORI dans nnUNet, un module de Transformer Swin a été implémenté. Ce module a été inséré après chaque couche de convolution de l'encodeur nnUNet, permettant à LORI d'utiliser les cartes de caractéristiques générées par les convolutions.

Les résultats expérimentaux, présentés dans un tableau spécifique, mettent en évidence la supériorité de la méthode proposée, LORI, par rapport aux approches de pointe sur trois ensembles de données divers. L'évaluation démontre l'efficacité notable de LORI dans la segmentation précise de multiples organes dans une image CT 3D. En particulier, LORI réalise une amélioration significative du score de Dice, avec +1.6 points sur l'ensemble de données WORD et +0.35 points sur l'ensemble BCV par rapport à la deuxième meilleure méthode. De plus, LORI montre une amélioration remarquable de +0.86 points sur l'ensemble de données LIVUS, caractérisé par des images échographiques très bruitées. Ce résultat souligne la capacité de LORI à améliorer la qualité de la segmentation dans des modalités difficiles, établissant ainsi son efficacité pour relever des tâches de segmentation complexes. De plus, en considérant la distance moyenne de Hausdorff à 95%, LORI atteint une valeur de 6.45 comparée à 7.90 pour la deuxième meilleure méthode. La performance supérieure de LORI sur cette métrique souligne sa capacité à capturer avec précision les contours des organes et à produire des segmentations plus précises. L'évaluation de la distance moyenne symétrique de surface (ASSD) valide également la supériorité de LORI en termes de précision de segmentation. Avec un ASSD de 0.97mm par rapport à 1.86mm pour la deuxième meilleure méthode, LORI démontre un gain substantiel de 0.89. Ce résultat renforce l'idée que LORI surpasse les autres méthodes sur plusieurs métriques d'évaluation.

Conclusion et Perspectives

Dans cette thèse, le problème de la segmentation d'images de haute dimension, en particulier les images médicales 3D, a été abordé. La haute dimensionnalité de ces images représente un défi pour leur segmentation. Il a été expliqué que les modèles d'apprentissage profond (DL) nécessitent un contexte global pour segmenter efficacement les régions locales, tandis que les modèles classiques souffrent de champs réceptifs limités ou de tailles de régions d'entrée restreintes. Pour surmonter ces limitations, les modèles de Transformers ont été choisis pour leur capacité à capturer des interactions à longue portée. Nous avons développé des modules Transformers utilisant des tokens globaux pour améliorer la capacité des Transformers à modéliser des interactions longues et hors de portée sur des images de haute dimension, telles que les images médicales 3D.

L'utilisation des modèles et concepts présentés pourraient être une base pour plusieurs projets. Nous avons collaboré avec l'IRCAD sur une base de données publique comprenant des images couplées d'échographie et de CT du rein en 3D, spécialement conçues pour la segmentation et le recalage. Dans le cadre du processus d'évaluation, le modèle GLAM a été utilisé comme base pour la segmentation. L'IRCAD, avec le projet DISRUMPERE, vise à développer un dispositif médical pour la segmentation en temps réel des images échographiques, notamment pour la détection autonome de tumeurs, marquant une avancée importante en IA médicale, les global tokens permettraient de traîter de façon efficace l'aspect temporel de cette tâche. L'étude de Moor [41] souligne que les modèles de fondation, combinant texte et images, sont cruciaux pour l'analyse d'images médicales, grâce à leur traitement de données de haute dimension et leurs capacités multimodales, les tokens globaux pourraient être utilisés pour créer un échange d'information entre modalité. Enfin, les tokens globaux seraient utiles dans divers domaines de haute dimension avec des dépendances longues portées comme la vision par ordinateur en haute résolution et l'analyse audio.

Chapter 1

Introduction

Contents

1.1 Context **38**
1.2 Motivations and challenges **41**
1.3 Main trends in medical image segmentation **45**
1.3.1 Medical image segmentation before deep learning 45
1.3.2 Convolutional neural networks (CNNs) 45
1.3.3 Transformers 47
1.4 Contributions and Outline **54**
1.5 Related publications **57**

1.1 Context

Artificial intelligence (AI) has emerged as a prominent and rapidly evolving field of research, garnering widespread attention and participation from researchers across disciplines. Notably, deep learning (DL), as a major branch of AI, has witnessed significant advancements, resulting in accelerated progress. Multiple historical tasks have benefited from this progress. Fig 1.1 illustrates some of these tasks with image and video analysis, audio recognition, or text translation. This notable leap forward has been made possible due to two main factors: First, the development of extensive datasets, such as ImageNet [7] with 14M images and 21K classes for image classification, ADE20K [31] with 30K images and 150 classes in image segmentation or COCO [42] with 330K images and 80 classes for object detection. Secondly, the increase in model parameters since 2012 was made possible by the development of graphics processing unit (GPU) computing which made DL model training much faster. More recently, novel techniques like Transformers [29] models have further contributed to leveraging these vast datasets effectively. Consequently, these advancements have facilitated the creation of models with a substantial number of parameters, commonly referred to as foundation models. This rapid development has arrived in the hands of the general public domain through the release of user-friendly software tools like Chat GPT [43, 44], which offer valuable assistance in various tasks such as translation, coding, and writing. Additionally, tools like DALL-E [45] and Midjourney enables the creation of realistic and artistic images based on textual prompts. Furthermore, there are ongoing efforts in numerous laboratories to develop applications of AI for the general public, such as autonomous driving and video games. Simultaneously, AI has begun to make significant inroads in specific fields such as climatology, astronomy, and medicine.

Computer vision (CV), as a subfield of AI, focuses on the processing and analysis of visual information from various sources such as cameras, images, and videos. CV includes tasks such as image classification, object detection, and image segmentation (Fig 1.2). Image classification involves the assignment of a specific label or class to an image. This task entails the identification and categorization of the main objects or features present within the image. Object detection, on the other hand, focuses on determining the precise position or location of one or multiple objects within an image. Object detection enables the identification of different objects and their corresponding spatial coordinates. Image segmentation is yet another crucial task in computer vision, which involves assigning a specific class to every individual pixel within an image. This process enables the partitioning of the image into distinct regions or segments, each associated with a particular category. Image segmentation plays a core role in medical image analysis.

The present thesis endeavors to enhance the field of medical image analysis with several CV research innovations. It is conducted in collaboration with the Conservatoire Nationale des Arts et

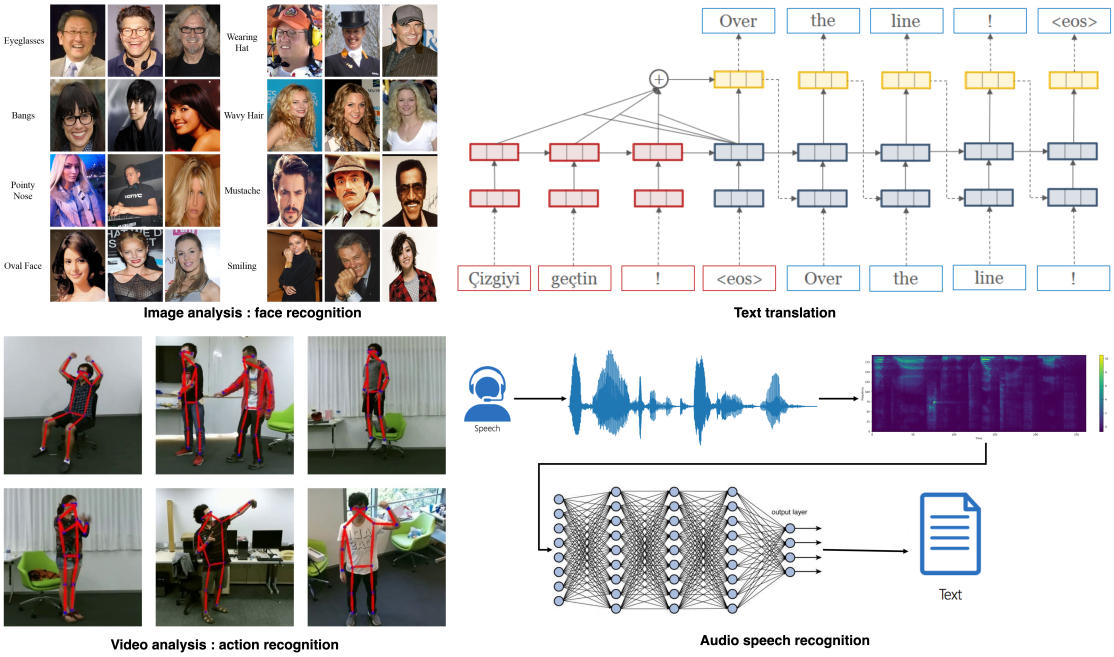


Figure 1.1: Significant applications of AI research: Faces recognition [3] and video Action recognition [4] in computer vision, Audio speech to text [5] for audio analysis and text translation [6] for natural language processing.

Figure 1.2: Examples of traditional computer vision tasks: Classification with an image of a Border Collie from ImageNet [7], Object detection with the detection of multiple objects (cars, bicycle, truck) in a city [8], and Segmentation of a city [9] with multiple objects segmented (tree, sky, cars, pedestrian).

Métiers (CNAM) in Paris and the Digestive System Cancer Research Institute (IRCAD) in Strasbourg. IRCAD is a renowned institution for medical innovation, globally recognized for its excellence in research. This thesis is part of the Disrumpere[1] project: Democratization of automatic diagnosis,

[1] https://www.ircad.fr/fr/newsletter-de-lircad-decembre-2022/

screening, biometrics and augmented percutaneous surgery assisted by artificial intelligence. The primary objective of Disrumpere is to increase access to healthcare by leveraging artificial intelligence to enhance the capabilities of affordable portable ultrasound probes, thereby making their utilization accessible for non-experts. Led by IRCAD France and IRCAD Africa, the project brings together teams of dedicated engineers and researchers who are determined to develop a tool that can significantly impact healthcare in Africa and address the challenges faced by "medical deserts" in France. The project encompasses several key components. The first component involves the development of high-performing algorithms for diagnostic and monitoring purposes. These algorithms aim to facilitate the easy detection of common pathologies, enabling efficient and accurate diagnosis. The second component focuses on the democratization of augmented percutaneous surgery through robotics. This aspect of the project involves utilizing ultrasound guidance to perform biopsies or destroy small cancerous tumors using needles. By making this surgical technique more accessible, Disrumpere aims to expand the reach of minimally invasive procedures and improve patient outcomes.

The application of AI in medical imaging is widely recognized as a field with immense potential to benefit society. AI has the capability to assist healthcare professionals in various tasks, ranging from diagnosis to surgical planning and guidance (Fig. 1.3). In the realm of medical image analysis, a diverse range of medical imaging modalities exists, each possessing its own benefits. These modalities encompass an array of images such as microscopic retina or cell images, Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and Ultrasound (US). MRI offers exceptional precision in capturing detailed information about soft tissues such as the brain, heart, or tumors without being invasive. CT, on the other hand, employs X-rays to obtain precise 2D or 3D scans of the entire body, including both tissues and bones. Ultrasound, a non-invasive and cost-effective modality, provides medical images in real time and without the harm of ionizing radiation. With the availability of these various imaging modalities, numerous tasks can be accomplished using AI [22, 23, 1, 24], including tumor or lesion detection, image segmentation, image reconstruction, and organ segmentation for diagnosis help or surgery planning.

Semantic segmentation has always been a fundamental task in medical image analysis [46, 47], as it is usually the first step in the chain of computer-assisted medical diagnosis. It can be used as a tool for clinical diagnosis as well as for surgery preparation and assistance. The main objective of this thesis is to address the task of segmenting medical images through the development of novel DL models. The focus of this study is on two modalities, namely CT and US images. The segmentation of CT scans is of great significance due to the existence of numerous applications, datasets, and research efforts dedicated to this task. Integrating the segmentation of CT volumes into the Disrumpere project for medical image registration between Ultrasound and CT holds immense potential. Moreover, US image segmentation is of utmost importance, as manual segmentation of ultrasound images by experts

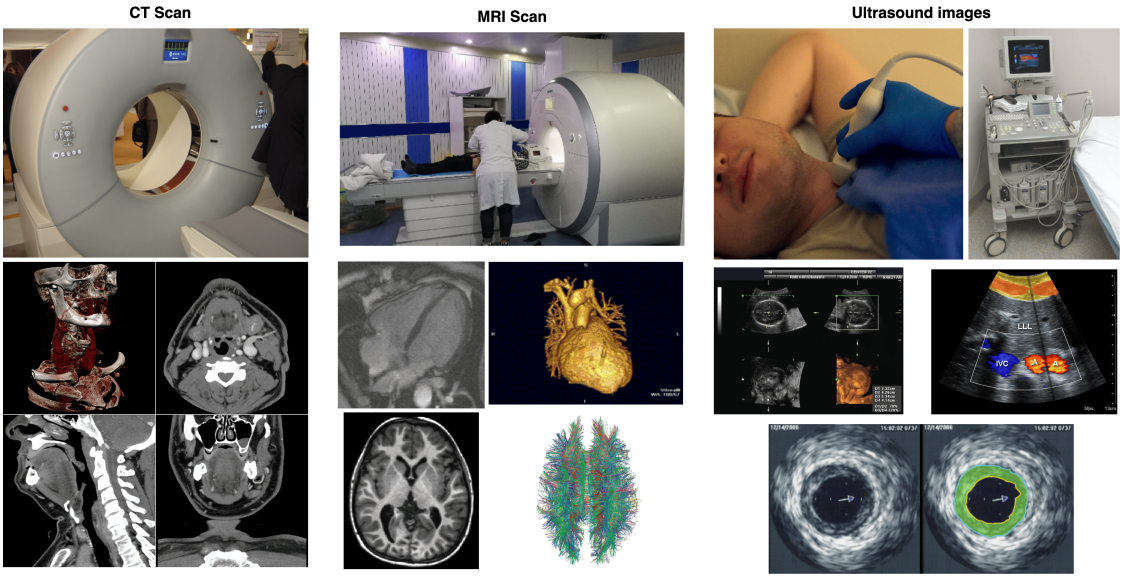


Figure 1.3: Example of medical image analysis modalities (CT: Computed Tomography; MRI: Magnetic Resonance Imaging; US: Ultrasound images).

can be time-consuming and challenging. Using ultrasound images for medical purposes provides several advantages. Firstly, ultrasound images can be obtained quickly, allowing for timely diagnosis and treatment. Secondly, the use of ultrasound is less invasive compared to other imaging techniques, reducing patient discomfort and potential complications. Additionally, ultrasound imaging is a cost-effective and real-time option, making it more accessible for medical practitioners and patients alike. Furthermore, the 3D aspect of medical images segmentation is of significant importance. In CT, it helps to detect small and difficult organs by giving more contextual information. But this importance is greater with US images. Unlike 2D ultrasound images, which lack sufficient information for accurate segmentation, 3D ultrasound images (Fig. 1.4) provide a greater level of detail. This enhanced level of information enables more precise and reliable automatic segmentation, enhancing the diagnostic capabilities of ultrasound imaging. Automating this process through advanced DL techniques has the potential to provide the medical community with new tools. The integration of this thesis with other research endeavors within the Disrumpere project opens up exciting prospects for research in the field of accelerated and non-invasive automated surgical procedures.

1.2 Motivations and challenges

Ultrasound image segmentation is a complex task that is faced with significant challenges, primarily due to two key factors: the limited availability of annotated data and the inherent quality associated

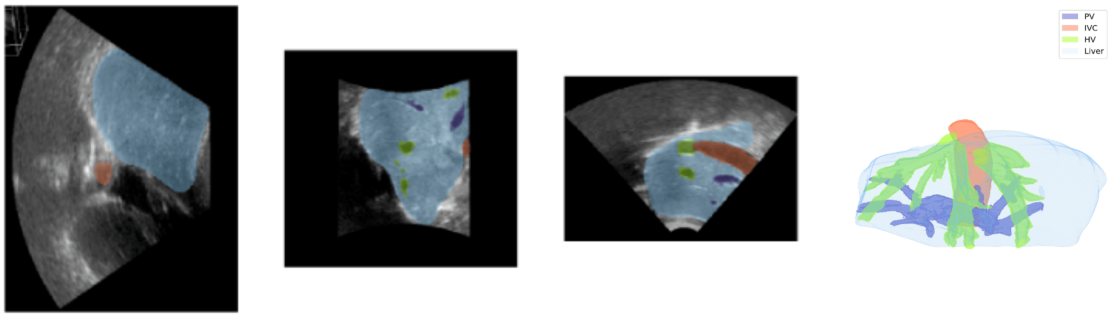


Figure 1.4: 3D Ultrasound image and segmentation of Liver and Vessels. The segmentation mask is done by an expert. These images come from a private dataset collected at IRCAD.

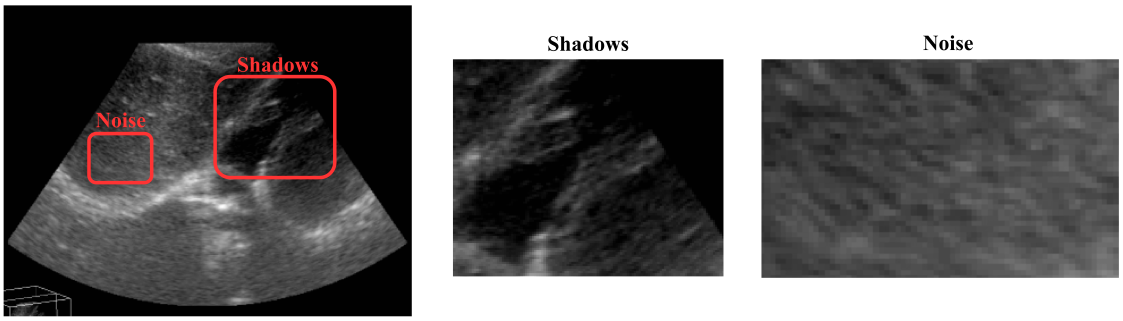


Figure 1.5: Challenges in US images interpretation: US images show multiple sources of data uncertainty. As shown in the image, there are often acoustic shadows which hide parts of the image. Compared to CT and MR, US usually has a higher signal-to-noise ratio, which reduces the capacity to find details as shown on the image's zoom.

with US modality. As shown in 1.5, US images are inherently noisy, with speckle noise posing a particular difficulty in the segmentation process. Furthermore, the resolution of US images varies within the volume, leading to distortions in the image. Additionally, certain tissues have the ability to completely absorb the ultrasound wave, resulting in occlusions and shadows within the images. These challenges impede the effective use of conventional deep learning models for US image segmentation. These artifacts pose difficulties even for expert clinicians, as evidenced by deviations in segmentation outcomes among different experts, as demonstrated in Fig. 1.6. As a result, traditional deep learning models encounter limitations when applied to US modalities, with errors on edges and confusion on difficult tasks like the segmentation of multiple types of vessels [48, 49]. Consequently, there is a growing interest in strategies that leverage contextual information to overcome these challenges in the research community. By incorporating additional context, the model's ability to accurately classify a noisy or occluded pixel can be enhanced by considering long-range dependencies. In bigger noisy or shady regions, the model must extend its analysis beyond the immediate neighborhood in order to discern the prevailing structures within said region.

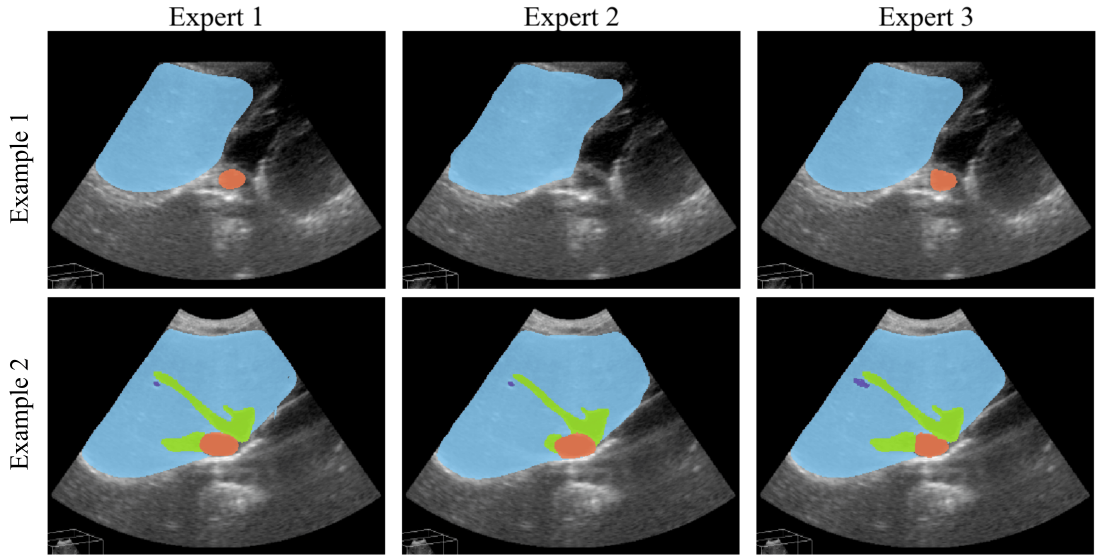


Figure 1.6: Two examples of Ultrasound images with Liver and vessel segmentation by three experts. We see on the first row the disagreement to segment the IVC (in orange) with Expert 2 vs Expert 1 & 3. On the second row, there is a disagreement about segmenting the HV (in green) with Expert 2 vs Expert 1 & 3, and the liver (in blue) with Expert 3 vs. Expert 1 & 2.

Medical imaging has significantly benefited from the use of 3D images, which allow for more comprehensive context modeling. Utilizing volumes instead of 2D slices is a superior approach as it provides more information to the model during voxel segmentation. With the ability to examine structures from multiple directions, the model can better leverage the context in US, CT, or MRI segmentation. While multiple architectures that directly exploit volumes exist, they often require a high spatial complexity, resulting in a significant consumption of GPU memory. When employing a state-of-the-art method such as 3D-UNet [11], it's impossible to process the entire volume of the image. It is common for 3D images to have large dimensions. However, even utilizing only the first layer of this 3D image would consume approximately 24 GB of GPU memory. Moreover, the entire model would necessitate more than 60Gb of GPU memory for processing a single volume during the inference stage, and over 120Gb during the training stage. This poses a significant challenge due to the high cost of GPUs, which are expensive tools used for DL model training. As a result, this hardware limitation can have a detrimental impact on the capacity and capability of DL models. To address this problem, researchers have adopted a common strategy of training DL models on a smaller region of the original volume [1, 25], allowing the model to process smaller portions of the volume during both training and inference. This method enables the model to segment the full volume using a sliding window strategy during inference. But the common strategy is to work with cropped volume patches. A typical configuration would be for a medical dataset to show an average volume size of $512 \times 512 \times 256$

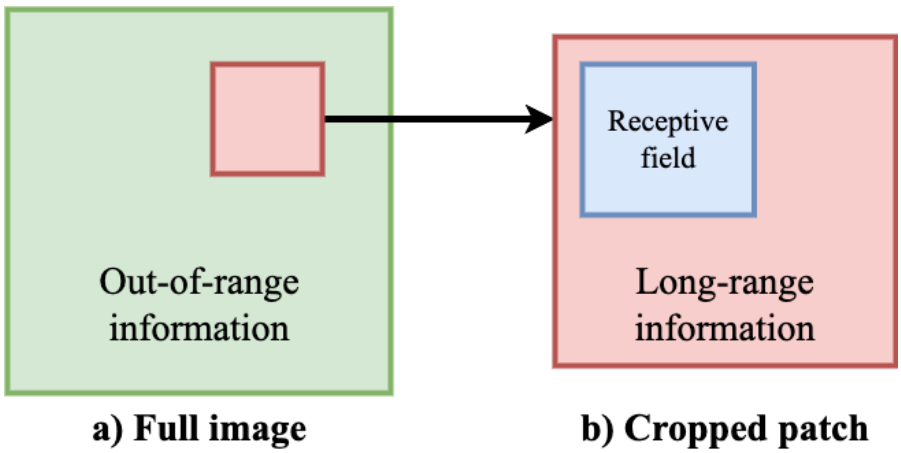


Figure 1.7: **Long and out-of-range limitations.** This schema shows the out-of-range information loss effect due to random cropping, a commonly employed strategy for training segmentation models on large 3D medical images. On b) we see a zoom on the cropped patch (in red) from the original image (in green). The Receptive Field (in blue) in a CNN designates the area of the input image reached by a unit at the end of the network. When classifying the pixel in the center of the blue square, the area outside the receptive field is thus not taken into account. The information outside the patch in the original image is, by definition, not used during segmentation. Through this manuscript, we designate as long-range the information not encompassed by standard CNN backbones receptive field and as out-of-range areas outside the cropped patch.

voxels and to train on random crops of dimension $128 \times 128 \times 64$, which represents only 1.6% of the original volume. This implies a loss of the total information and context available for the DL model to exploit. As can be seen in Fig. 1.7.a), the cropping strategy consists of selecting a smaller patch in a full image to be processed by the model. Because of this strategy, the methods lose what we call out-of-range information, information outside of the red-cropped image. Furthermore, even within the cropped patches, most methods can't model interactions long-range interactions specifically within high-resolution features. On 1.7.b), we denote as long-range information, the information outside the model receptive field in blue.

With the awareness of all aforementioned limitations, the objective of this thesis is to investigate new methodologies to improve the accuracy of 3D medical image segmentation. **This thesis aims to explore approaches that effectively use global context in 3D images, which means modeling interaction inside and outside of the cropped patches, while simultaneously maintaining a high spatial resolution to leverage more information, and consequently, to achieve more accurate segmentation.**

1.3 Main trends in medical image segmentation

1.3.1 Medical image segmentation before deep learning

The automatic segmentation of objects in medical images has been a topic of significant interest and extensively explored in the literature [50, 51, 52]. The initial attempts primarily focused on hand-crafted methods, where segmentation relied on the image itself and predefined rules such as thresholding [53, 54], region-growing [55, 56], or watershed [57, 58]. Subsequently, model-based approaches gained prominence with the increasing availability of labeled data. These approaches encompass deformable models [59, 60, 61, 62] and atlas-based methods [63, 64, 65, 66, 67, 68]. Deformable models aim to modify a predefined curve to accurately fit the image through the utilization of energy minimization algorithms. On the other hand, atlas-based methods are more specific to medical images and exploit the fact that organs are typically located at similar positions across patients. These methods utilize label-transfer to assign labels to the target volume based on annotations from the dataset. Additionally, Statistical Shape Models (SSMs) [69, 70, 71] are often employed to constrain the models with shape information extracted from labeled images.

1.3.2 Convolutional neural networks (CNNs)

Convolution [72] is a fundamental image processing method involving the application of a weighted filter to an image, enabling operations like edge detection by considering the surrounding pixels' influence. It should be noted that convolution is a local operation, meaning that it applies its weights to a small region of the input, but shares the same weights for the whole input. Consequently, the number of parameters utilized in convolutional operations does not depend on the size of the image and can be easily controlled. Multiple convolutional neural network (CNN) architectures have been developed and refined over the years, leveraging the inherent power of convolutions. These architectures have brought about a significant revolution in computer vision, primarily in image classification, but also extending to semantic segmentation tasks, but also to signal processing field, time series or audio. The field of image classification with deep learning witnessed its initial breakthrough with the introduction of AlexNet [10] (see Fig. 1.8), which incorporated a hierarchical structure of convolution layers. This was followed by the development of VGG [73], which featured a deeper architecture with smaller convolution kernels. Subsequently, Inception [74] introduced skip connections and multiple convolution kernel sizes to enhance performance. Finally, residual neural networks (ResNet) emerged as a significant advancement in deep learning for computer vision [75]. This architecture introduced a residual layer that facilitates identity mappings and enables deep learning models with tens or hundreds of layers to train easily while achieving improved accuracy when going deeper. Residual Networks

have become a crucial component of deep learning models in computer vision.

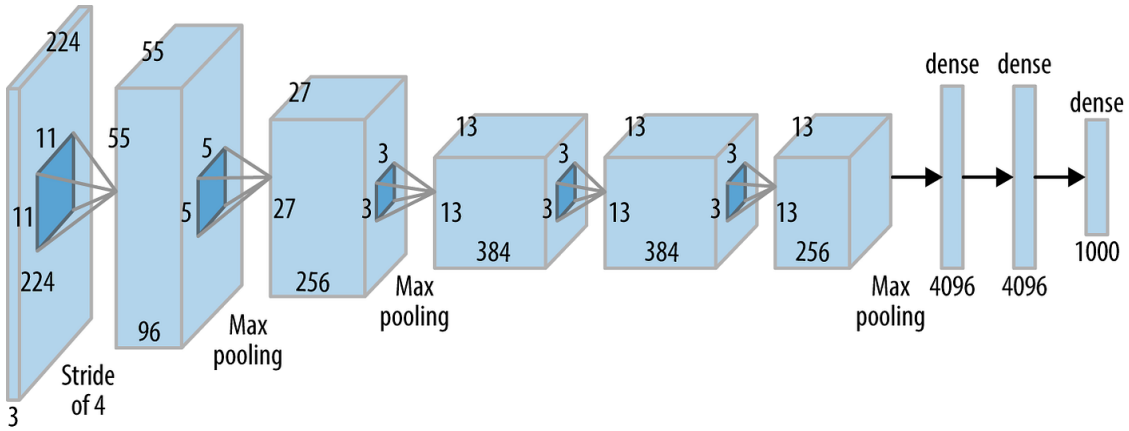


Figure 1.8: **AlexNet schema [10].** Simplified schema of AlexNet with the different succession convolution and pooling layers, followed by dense layers.

In the current era of deep learning, Fully Convolutional Neural Networks (FCNs) [76, 77, 78, 79, 80] have been at the forefront of state-of-the-art performance in semantic segmentation. In this field, the approach typically involves separate encoder or backbone networks, with the aim of extracting maximum semantic features from the images [75, 81, 82, 83]. The decoder network, on the other hand, is responsible for constructing the segmentation mask [84, 85]. For instance, DeepLab [79] is a well-known model that is based on an encoder-decoder architecture. Various datasets are used by the research community to adapt and improve the performance of these segmentation architectures [9, 86, 87, 88].

CNNs for medical images segmentation. Convolutional neural networks have emerged as powerful tools in the field of medical image analysis due to their ability to learn and extract relevant features, implemented as convolutional filters, from images. Following the ideas of FCNs, medical segmentation methods are FCNs-based methods with skip connections, an innovation well suited for medical images because of the structured data aspect of medical images. Conventional architectures such as UNet [20] and its variants, purpose-built for medical image segmentation, are commonly employed. Notably, UNet's architecture, which incorporates skip connections between encoder and decoder modules, is well-suited for small datasets when associated with a strong data augmentation as it facilitates the integration of high-resolution features from the encoder into the decoder, enabling effective utilization of input image information. UNet [20] and 3D UNet [89] architectures, which are CNN based presented in Fig. 1.9, have demonstrated their effectiveness in various medical image segmentation tasks. Furthermore, the multi-scale structure of UNet models enables them to capture local details in high-resolution feature maps and extract increasingly higher-level semantic features in each down layer until the bottleneck. However, despite their impressive performance, these models

are limited by their receptive field which refers to the area in the input image that a particular feature in the network can "see" and use to make a prediction [18], this is called the receptive field in a CNN and it designates the area of the input image reached by a unit at the end of the network. In other words, it represents the effective size of the convolutional kernel at a given layer. The receptive field of a feature is determined by the size and number of layers in the network, as well as the stride and pooling operations applied. This limitation is particularly pronounced in high-resolution feature maps, where the receptive field can be small and the amount of information available for segmenting a pixel is limited due to the local nature of the convolutions.

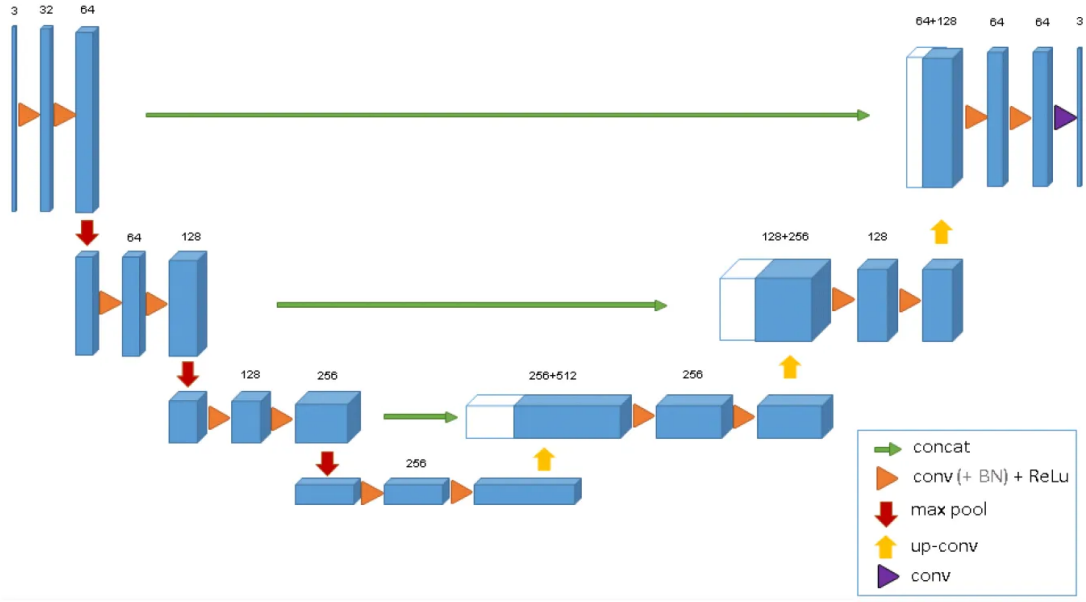


Figure 1.9: 3D UNet architecture schema [11].

1.3.3 Transformers

Transformers in NLP. Transformers have emerged as a recent breakthrough in NLP, revolutionizing the way language models are designed and trained [90]. They incorporate a novel combination of multi-head self-attention, multi-layer perceptron, skip connection, and layer normalization to handle long sequences and capture long-range dependencies. The self-attention mechanism connects each input element, denoted as tokens, to each other which is well designed to model long-range interactions. This is a significant improvement over previous recurrent models [91, 92], which struggled with long-range dependencies. The introduction of transformers enabled the development of powerful models such as BERT [93], which outperformed every prior NLP model. Subsequently, a variety of

large-scale Transformers-based models such as GPT [43, 44, 94] have been introduced, leading to a revolution in the NLP field.

The Transformers self-attention mechanism is a critical component of the architecture. To elaborate, the mechanism works by projecting a sequence of N embedded tokens, $X \in \mathbb{R}^{N\times d}$, where d denotes the embedding dimension, into queries $Q \in \mathbb{R}^{N\times d}$, keys $K \in \mathbb{R}^{N\times d}$, and values $V \in \mathbb{R}^{N\times d}$. Subsequently, the attention is computed between the queries and the keys using the softmax function with the formula:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)$$

where $A \in \mathbb{R}^{N\times N}$. The output of the self-attention mechanism is then obtained by computing $Z = AV$, where $Z \in \mathbb{R}^{N\times d}$. This output can be interpreted as a weighted sum of the values V with the weights given by the attention scores in A. The self-attention mechanism thus enables the Transformers to capture dependencies between tokens at various positions in the input sequence, providing a powerful tool for modeling long-range dependencies. The original transformer model is detailed in Fig. 1.10.

Transformers in computer vision. The pioneering work before the use of Transformers in computer vision can be traced back to the field of video analysis, as demonstrated in the work by [95]. In this work, self-attention mechanisms were employed to effectively model long-range dependencies between image frames. Transformers have gained popularity in the computer vision community due to their ability to model long-range dependencies and perform full attention on input data. The Vision Transformers (ViT) [12] is the first to employ a Transformer encoder directly for image classification and has demonstrated superior performance compared to other methods. ViT has demonstrated that Transformers possess the capability to effectively handle extensive datasets when trained on ImageNet, resulting in great performance outcomes. To adapt transformers for image processing, ViT divides images into patches of size 16×16 and re-embeds them into a sequence of N tokens of size d. A class token is added to the sequence as an extra learnable token, which serves as the features for the final classification layer. This class token inspired us in our work for the development of the global tokens which will be explained later. Additionally, ViT uses positional encoding to bring positional bias to the tokens, which can take various forms [96, 97, 98] (see Fig. 1.11). Although ViT is a strong backbone for various tasks, it is challenging to train [99] as transformers require more data than CNNs because it has many more trainable parameters, leading to a difficult pre-training phase. Moreover, ViT computes a full attention map between all its input tokens, resulting in spatial complexity of $\mathcal{O}(N^2)$ per layer. This large complexity could be problematic for high-dimensional images because of memory consumption limitation and it limits the possibility of working with smaller patches for finer-grained information modeling. Some methods like windowed transformers in hierarchical models [17] exist to tackle these issues and will be described in Chapter 3.

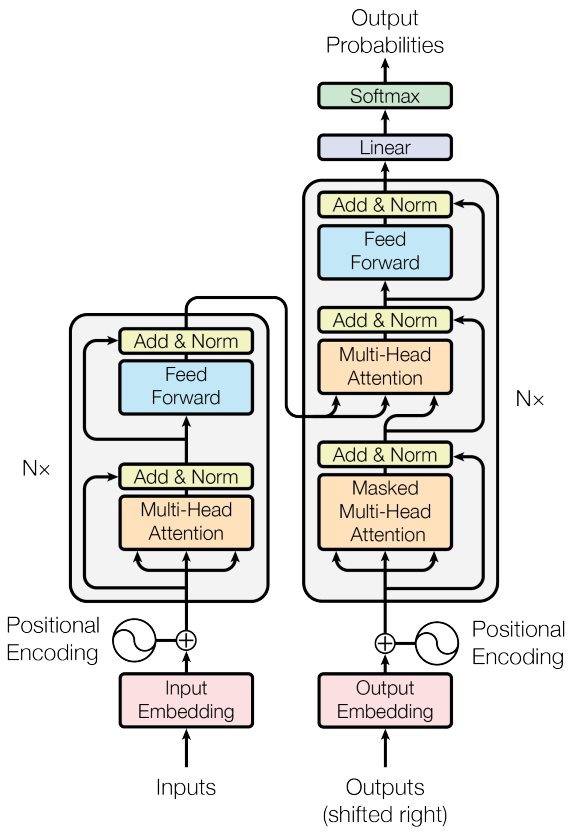


Figure 1.10: **The transformer architecture.** This schema shows the detailed original transformer model architecture with the encoder and decoder parts and all its sub-modules: embedding of the input tokens, positional encoding, multi-head self-attention, normalization, skip connections, and feed-forward network.

More recently, the Segment Anything Model (SAM)[100] was developed using transformers, which has demonstrated remarkable ability to segment all parts of an image without the need for labels. This architecture is able to segment by looking for semantically coherent components related to a pixel. However, Transformers still face challenges in the domain of medical imaging, particularly in ultrasound medical images, as shown in 1.12 mainly because of the high dimensionality of medical images and domain shift.

Transformers for medical image segmentation.

The medical imaging community has been actively engaged in researching attention mechanisms [101, 102, 103, 104, 105, 19, 106]. Among these models, Attention U-Net [19] introduces an additive attention gate to selectively filter the features obtained from the skip connections. In this thesis, we investigate the attention mechanism introduced by Transformers, which incorporates a self-attention mechanism enabling connections between all input entries. This feature greatly facilitates the

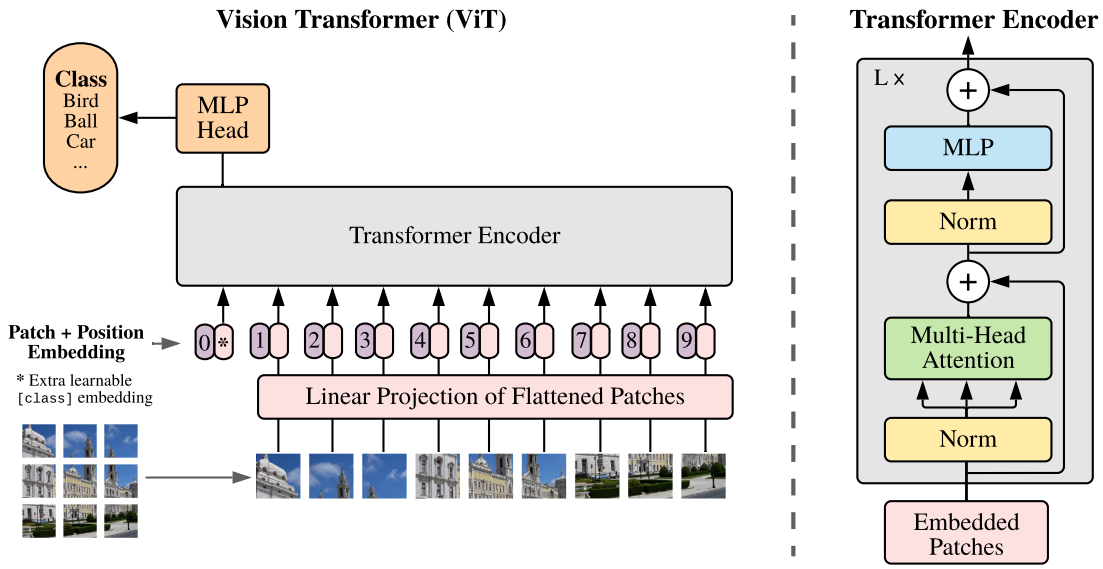


Figure 1.11: **ViT model schema [12].** ViT splits an image into fixed-size patches, linearly embeds each of them, add position embeddings, and feeds the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, ViT use the standard approach of adding an extra learnable "class token" to the sequence.

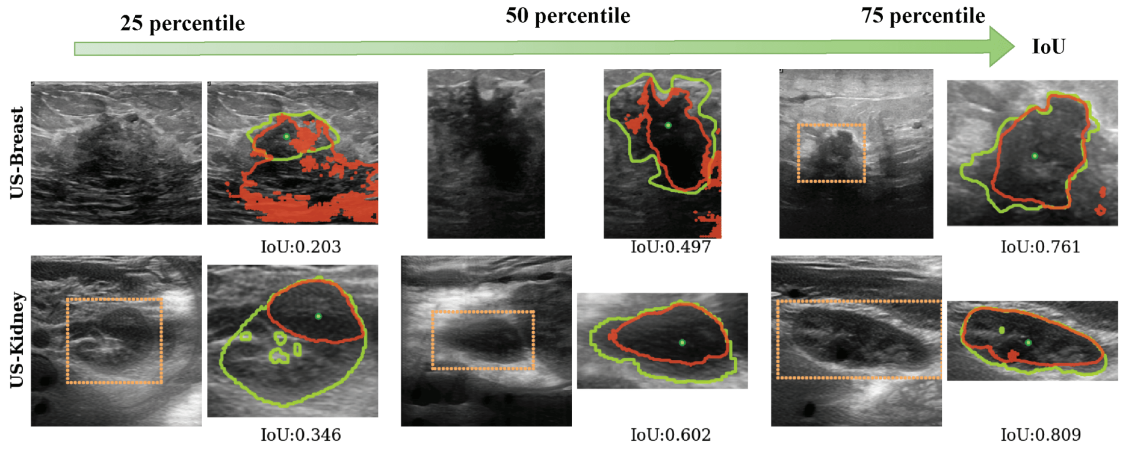


Figure 1.12: Example of segmentation produced by SAM on US images [13]. We observe that SAM struggle to segment the breast tumour and the kidney. Even if SAM is trained on an extra large dataset and is design to be an universal segmentation model, it suffers from domain shift.

modeling of global context. Transformers have significantly benefited the medical image community by introducing robust models for image segmentation [36, 107, 14, 35]. Each of these approaches leverages transformers to enhance the receptive field of the method, thereby leveraging additional information from the input to generate the segmentation results. TransUNet [107] is a pioneering technique for segmenting medical images using Transformers. This method utilizes a hybrid encoder

architecture that combines CNNs and Transformers to extract essential information from the input image. Subsequently, a CNN decoder is employed to generate the segmentation mask. It is important to note that TransUNet is designed exclusively for processing 2D images, which poses a significant limitation as it cannot be applied in scenarios where a 3D context is required. UNETR [36] employs a transformer-based encoder architecture, similar to the ViT [12], whereby the entire 3D input is subdivided into smaller patches, which are then treated as tokens and processed within the Transformer modules. Subsequently, the extracted features undergo recombination and further processing through a CNN decoder. While UNETR serves as a strong model, it is noteworthy that its encoder component demands a substantial amount of GPU memory. Consequently, harnessing UNETR to its full potential necessitates substantial computational resources, approximately eight times more than what is typically required by other comparable methods.

Moreover CoTr [14], as illustrated in Fig. 1.13 represents a hybrid model comprising both a CNN encoder and decoder components. Notably, CoTr introduces a deformable transformer module positioned between these encoder and decoder segments, employed to amplify feature extraction capabilities. The deformable characteristics inherent to this transformer module afford CoTr the ability to approximate self-attention mechanisms when handling voluminous 3D input images. However, it is imperative to acknowledge that CoTr is subject to certain limitations. Specifically, the self-attention mechanism within CoTr does not encompass the high-resolution feature maps generated by the encoder. Lastly, nnFormer [35] (or Swin-UNet) represents a transformer-based model comprising both an encoder and decoder fashioned entirely with transformer components. This model stands as a robust baseline, capitalizing on windowed Transformers [17], which effectively approximate the self-attention mechanism by computing it locally, as opposed to considering the entirety of the input. Additionally, nnFormer leverages a hierarchical architectural design, facilitating feature mixing across the model's layers. It is worth noting that the window transformer paradigm employed in nnFormer demonstrates substantial proficiency in processing high-dimensional images. However, it is important to acknowledge a trade-off in this approach. While it excels in handling large-scale image data, it does exhibit limitations in terms of global context modeling, particularly within high-resolution feature maps.

Efficient attention in Transformers.

Long sequences have been a challenge for transformers because the original self-attention mechanism has a quadratic complexity in the sequence length. Thus, efficient attention mechanisms have garnered significant attention in recent years as shown on Fig.1.14. The quest for more computationally and memory-efficient models has led to a proliferation of novel approaches and techniques. Here, we provide an overview of key developments in the field of efficient attention, categorizing them into several major themes and highlighting notable models and methods. It is noticeable that none of the methods are specified for 3D medical images which implies specific issues to deal with.

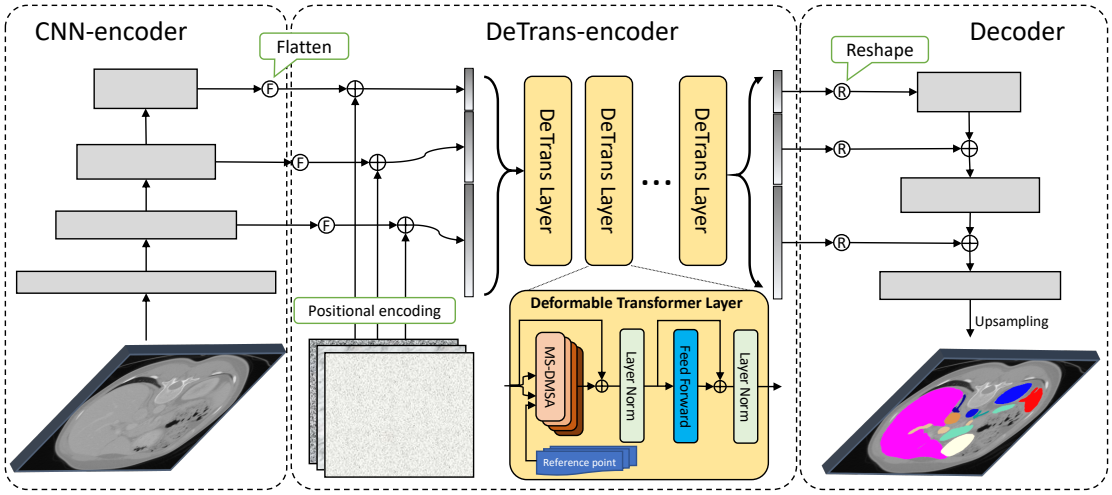


Figure 1.13: **Schema of CoTr [14].** A CNN-encoder, a DeTrans-encoder, and a decoder. Gray rectangles: CNN blocks. Yellow rectangles: 3D deformable Transformer layers. The CNN-encoder extracts multi-scale feature maps from an input image. The DeTrans-encoder processes the flattened multi-scale feature maps that embedded with the positional encoding in a sequence-to-sequence manner. The features with long-range dependency are generated by the DeTrans-encoder and fed to the decoder for segmentation.

A significant portion of research on efficient attention models falls under the category of sparsity and pattern-based approaches. These methods aim to reduce the quadratic complexity of self-attention mechanisms. Notable models in this category include Sparse Transformers [108], which introduced the concept of structured sparsity patterns in self-attention. They allow for selective attention to specific tokens while ignoring others, enabling a significant reduction in computational requirements. Other models like Routing Transformers [109] and Reformers [110] focus on learning adaptive patterns for attention, improving the scalability of the Transformer architecture.

Efforts to approximate self-attention mechanisms using low-rank approximations and kernel methods have emerged as a prominent trend. Models like Linformer [16], illustrated on Fig.1.15, propose a fixed low-rank factorization of the attention matrix, significantly reducing computational complexity. Performer [111] introduced a kernel-based approach to efficiently approximate self-attention, emphasizing scalability and the ability to handle long sequences.

In this thesis, we focus on high-resolution and structured data. Thus we stand in another method family based on computer vision strategies. Among these approaches, window-based patch extraction vision transformers recently provided a simple yet efficient approach to compute attention [17, 112, 113] (see Fig. 1.16). Some vision transformers have combined multiple efficient attention mechanisms. The recent ViT-inspired backbone PvT [114] is based on windowed self-attention and attention

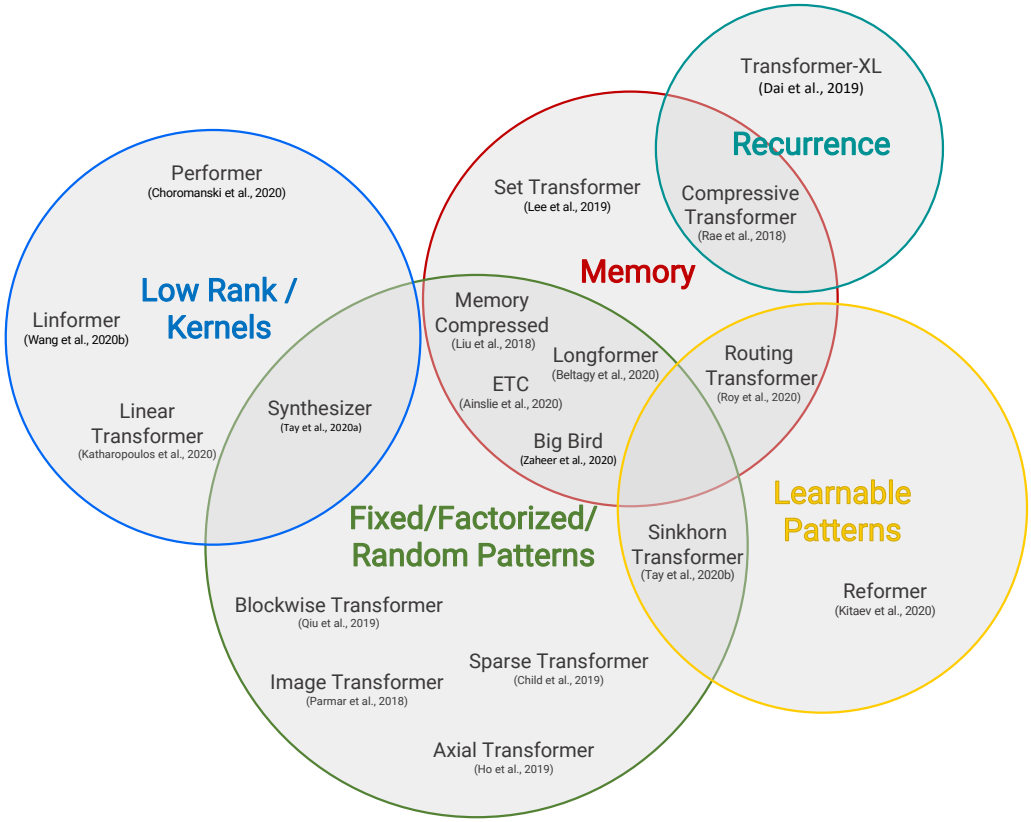


Figure 1.14: Taxonomy of Efficient Transformer Architectures [15].

approximation close to Linformer [16].

In the pursuit of memory efficiency, a class of models has been developed, incorporating global memory elements [115, 116]. Longformer [117] and ETC [118] are notable examples that include global memory components, allowing information to be shared across tokens more efficiently. ViL [119] balances sparse attention by using a reduced set of global tokens (usually a single one) to extract global representations of the input image. These models offer solutions for tasks that require processing long sequences, such as document summarization and question answering.

The aforementioned strategies have been primarily devised for tasks related to NLP or image classification. Consequently, these strategies are not specifically tailored for segmentation tasks since they do not include the preservation of attention across high-resolution feature maps. Additionally, none of these methods address the problem of incomplete representation of the entire volume in the input due to the usage of cropped patches.

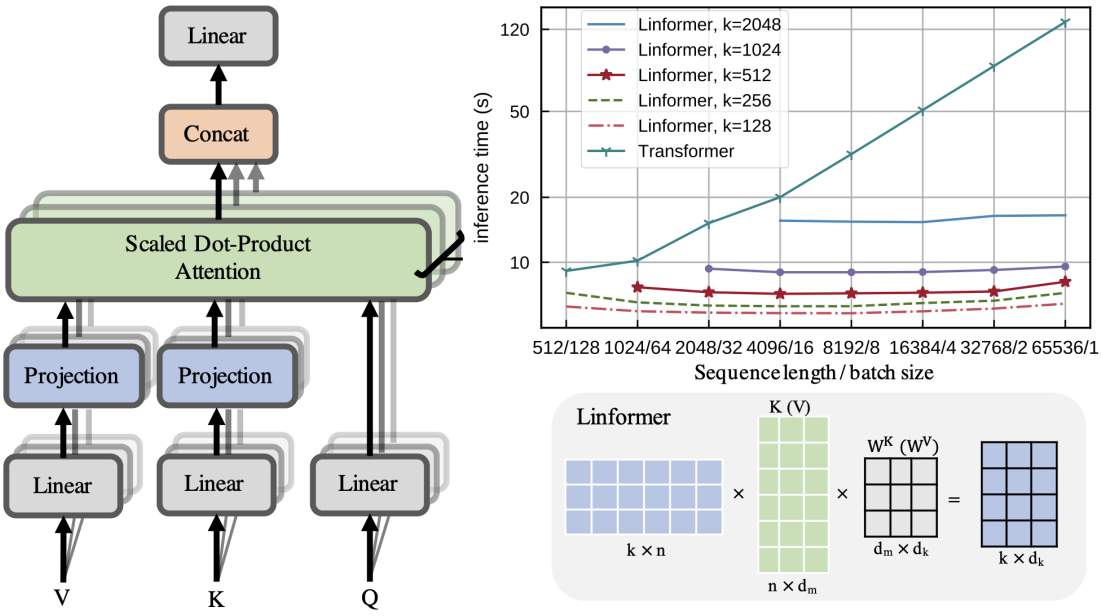


Figure 1.15: **Linformer [16].** Left and bottom-right show architecture and example of Linformer multihead linear self-attention. Top right shows inference time vs. sequence length for various Linformer models.

1.4 Contributions and Outline

Chapter 2: U-Net Transformer: Self and Cross Attention for Medical Image Segmentation

We first incorporated Transformers attention mechanism in a UNet model to address the problem of limited receptive field encountered in UNet architectures when applied to the task of segmenting 2D medical images. We introduce the U-Transformer network, which combines a U-shaped architecture for 2D medical images segmentation with self- and cross-attention from Transformers. U-Transformer was among the early adopters of Transformers for medical image segmentation. It enables modeling global attention in the bottleneck of the encoder, in contrast to standard attentions models used in the field (*e.g.* Attention-UNet [19]) that do not improve global context. Additionally, we demonstrated one of the earliest applications of Transformers in computer vision, following the publication of ViT [12]. U-Transformer overcomes the inability of U-Nets to model long-range contextual interactions and spatial dependencies, which are arguably crucial for accurate segmentation in challenging contexts. To this end, attention mechanisms are incorporated at two main levels: a self-attention module leverages global interactions between encoder features, while cross-attention in the skip connections allows a fine spatial recovery in the U-Net decoder by filtering out non-semantic features. Experiments show the large performance gain brought out by U-Transformer compared to U-Net and local Attention U-Nets. We also highlight the importance of using both self- and cross-attention, and the nice interpretability

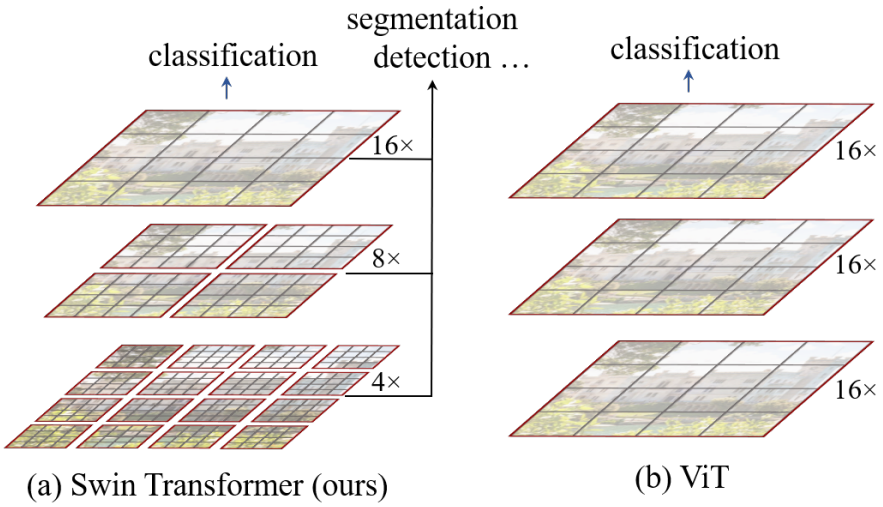


Figure 1.16: **Comparison between windowed based transformer and ViT [17].** (a) Swin Transformer builds hierarchical feature maps by merging image patches (shown in gray) in deeper layers and has linear computation complexity to input image size due to computation of self-attention only within each local window (shown in red). It can thus serve as a general-purpose backbone for both image classification and dense recognition tasks. (b) In contrast, previous ViT-produced feature maps of a single low resolution and have quadratic computation complexity to input image size due to computation of self-attention globally.

features brought out by U-Transformer.

Chapter 3: Full Contextual Attention for Multi-resolution Transformers in Semantic Segmentation

U-Transformer struggles to handle the high dimensionality of 3D medical images due to the quadratic complexity of self and cross attention, leading to limited efficiency in modeling long-range interactions with high-resolution features. The second phase of this thesis aimed to address the challenge of incorporating long-range interaction modeling into transformer models for both high-resolution images and 3D images, particularly at feature maps with high-resolutions. We present GLAM (GLobal Attention Multi-resolution transformers) as a solution to address the limitation of multi-resolution transformers in capturing local interactions within high-resolution feature maps. GLAM is introduced as a means to overcome this problem and enhance the performance of transformers in handling high-dimension images. GLAM is a generic module that can be integrated into most existing Transformers backbones. GLAM includes learnable global tokens, which unlike previous methods can model interactions between all image regions, and extracts powerful representations during training. The incorporation of global tokens in the model, influenced by the class token [26, 12], serves as pivotal elements for the transmission of information across various regions of the image at each stage. Consequently, this integration of GLAM engenders extensive interactions over

long distances, which were previously absent in the model. Extensive experiments show that GLAM exhibit substantially better performances than the vanilla state-of-the-arts on 2D high dimensions images. Moreover, GLAM performs also well on 3D medical images.

Chapter 4: Long and Out of Range Interaction transformer module for 3D medical image segmentation

To go one step further, this chapter explores the possibility to model full contextual information including out-of-range interactions when training a model from local 3D patches. This approach restricts the ability to leverage information from the entire volume, beyond the boundaries of the input. Building upon the concept introduced by GLAM, which employs global tokens to indirectly model information, we extend this idea by incorporating out-of-range interaction modeling even in high-resolution feature maps. Notably, our contribution is, to the best of our knowledge, the first to integrate information beyond the cropped input volume in the context of 3D medical images segmentation. We introduce a method that aims to address the aforementioned challenges by facilitating the incorporation of long-and-out-of-range dependencies in medical segmentation models. This method incorporates global tokens and employs self-attention mechanisms to create long-range and out-of-range interactions. We provide two variant of this method: FINE (Full resolutIoN mEmory transformer), a fully transformer architecture which works as a preliminary proof of concept, and LORI which is a generic module allowing it to be seamlessly integrated into existing models such as nnUNet. We performed preliminary experiments on BCV with FINE showing its relevance, and extensive experimental evaluations with LORI on three distinct datasets: two 3D CT multi-organs segmentation datasets and one 3D ultrasound image dataset for liver and vessel segmentation. The results obtained from these evaluations demonstrate the substantial enhancement in segmentation performance achieved by LORI. Notably, LORI exhibited superior performance across multiple multi-class high-resolution 3D image datasets, irrespective of the different modalities involved.

1.5 Related publications

U-Transformer	Olivier Petit, Nicolas Thome, Clement Rambour, Loic Themyr, Toby Collins, Luc Soler. "U-Net Transformer: Self and Cross Attention for Medical Image Segmentation". Medical Image Computing and Computer Assisted Intervention (MICCAI), workshop Machine Learning in Medical Imaging (MLMI), Oral Presentation, 2021.	Chapter 2
GLAM	Loic Themyr, Clément Rambour, Nicolas Thome, Toby Collins, Alexandre Hostettler. "Full Contextual Attention for Multi-resolution Transformers in Semantic Segmentation". IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Oral Presentation, 2023	Chapter 3
FINE	Loic Themyr, Clément Rambour, Nicolas Thome, Toby Collins, Alexandre Hostettler. "Memory transformers for full context and high-resolution 3D Medical Segmentation". edical Image Computing and Computer Assisted Intervention (MICCAI), workshop Machine Learning in Medical Imaging (MLMI), Oral Presentation, 2022.	Chapter 4
LORI	Loic Themyr, Clement Rambour, Denis Coquenet, Nicolas Thome, Toby Collins, Alexandre Hostettler. "LORI: Long and Out of Range Interaction transformer module for 3D medical images segmentation", International society for optics and photonics, Journal of Medical Imaging (SPIE JMI), 2023, *Under review*	Chapter 4

Chapter 2

U-Net Transformer: Self and Cross Attention for Medical Image Segmentation

Contents

2.1 Introduction **61**
2.2 Related Work **63**
2.3 The U-Transformer Network **64**
2.3.1 Self-attention 65
2.3.2 Cross-attention 65
2.4 Experiments **66**
2.4.1 U-Transformer performances 67
2.4.2 U-Transformer analysis and properties 69
2.5 Conclusion **71**

Chapter summary

As explained in section 1.2, medical image segmentation remains particularly challenging for complex and low-contrast anatomical structures. In this chapter, we introduce the U-Transformer network, which combines a Transformers attention mechanism with a UNet architecture. U-Transformer overcomes the inability of U-Nets to model long-range contextual interactions and spatial dependencies, which are arguably crucial for accurate segmentation in challenging contexts. To this end, attention mechanisms are incorporated at two main levels: a self-attention module leverages global interactions between encoder features, while cross-attention in the skip connections allows a fine spatial recovery in the U-Net decoder by filtering out non-semantic features. Experiments on two abdominal CT-image

datasets show the large performance gain brought out by U-Transformer compared to U-Net and local Attention U-Nets. We also highlight the importance of using both self- and cross-attention and the nice interpretability features brought out by U-Transformer.

2.1 Introduction

Until recently, state-of-the-art methods rely on Fully Convolutional Networks (FCNs), such as U-Net and variants [20, 11, 27, 28]. U-Nets use an encoder-decoder architecture: the encoder extracts high-level semantic representations by using a cascade of convolutional layers, while the decoder leverages skip connections to re-use high-resolution feature maps from the encoder in order to recover lost spatial information from high-level representations.

Despite their outstanding performances, FCNs suffer from conceptual limitations in complex segmentation tasks, *e.g.* when dealing with local visual ambiguities and low contrast between organs. Structures exhibiting long-range spatial dependencies within regions characterized by low contrast can result in misclassification. Moreover, the classification of small organs or with significant shape variability, such as the Pancreas, poses additional challenges. This is illustrated in Fig 2.1a) for segmenting the blue cross region corresponding to the pancreas with U-Net: the limited Receptive Field (RF) framed in red does not capture sufficient contextual information, making the segmentation fail, see Fig 2.1c).

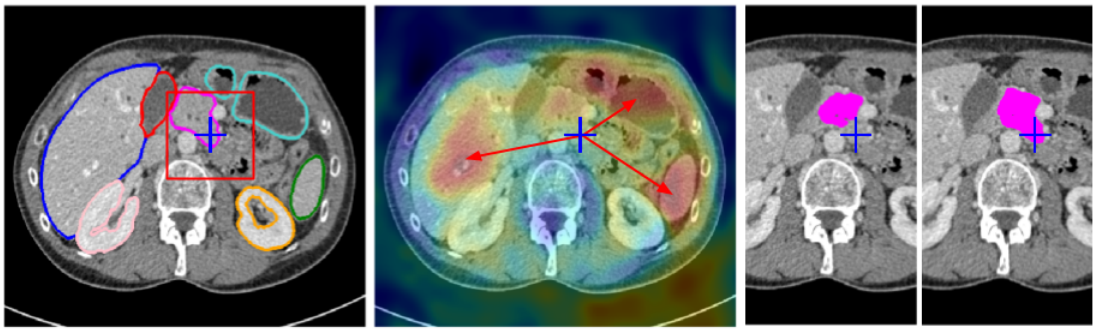
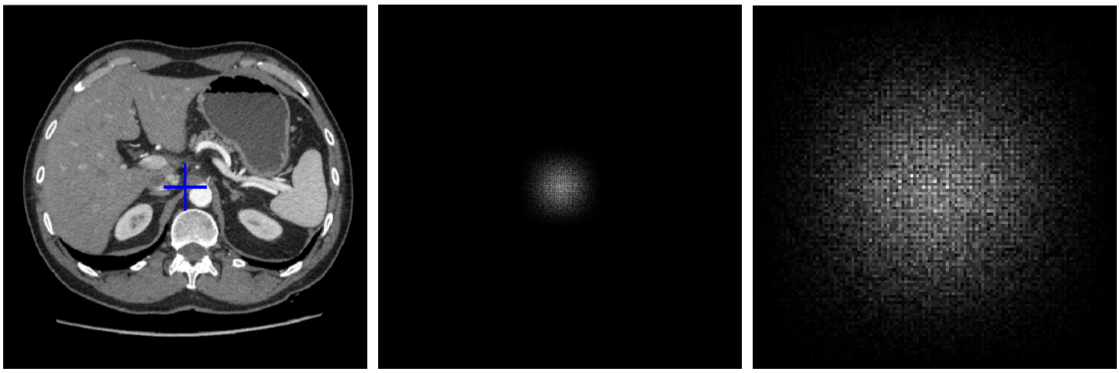


Figure 2.1: Global context is crucial for complex organ segmentation but cannot be captured by vanilla U-Nets with a limited receptive field, *i.e.* blue cross region in a) with failed segmentation in c). The proposed U-Transformer network represents full image context by means of attention maps b), which leverage long-range interactions with other anatomical structures to properly segment the complex pancreas region in d).

The Receptive Field (RF) in a ConvNet designates the area of the input image reached by a unit at the end of the network. It can be obtained theoretically by looking at the convolution and pooling operations. In our work, we use 512x512 input images and the Theoretical Receptive Field (TRF) of a standard U-Net is small (140x140) which does not enable to model full contextual information. Although the TRF is larger for deeper networks (e.g. nnU-Net), the TRF often overestimates the actual contextual information that the network could handle. This has been studied in [18] where the authors

introduced the notion of Effective Receptive Field (ERF). The proposed method consists of putting a gradient of one at the end of the bottleneck for the central unit and setting the other gradients to zeros. Then, we propagate the gradients with the back-propagation and get the values assigned to the network's input. Thus, we obtain an array with the same size as the input as shown in Fig. 2.2. We can see that the gradients describe a Gaussian centered on the image's center with the values quickly decreasing till reaching zero. With that Gaussian, we can get the Effective Receptive Field (ERF) as formulated in [18]. We can easily imagine that the ERF is considerably smaller than the TRF and for instance the nnU-Net, Fig. 2.2c, which has a large TRF gets a ERF of about 200x200, which is much lower.



(a) CT slice (b) U-Net (c) nnU-Net

Figure 2.2: **The Effective Receptive Field as formulated in [18].** We put a gradient of one at the end of the encoder and propagate it to the input. The figures show high gradient values in white and zero gradients in black. We analyze the U-Net and nnU-Net architectures and observe that the final ERF is much smaller than the TRF. a) b) and c) have the same dimensions (512x512 pixels).

In this chapter, we introduce the U-Transformer network, which leverages the strong abilities of transformers [29] to model long-range interactions and spatial relationships between anatomical structures. U-Transformer keeps the inductive bias of convolution by using a U-shaped architecture but introduces attention mechanisms at two main levels, which help model global context and interpret the model decision. Firstly, a self-attention module leverages global interactions between semantic features at the end of the encoder to explicitly model full contextual information. Secondly, we introduce cross-attention in the skip connections to filter out non-semantic features, allowing a fine spatial recovery in the U-Net decoder.

Fig 2.1b) shows a cross-attention map induced by U-Transformer, which highlights the most important regions for segmenting the blue cross region in Fig 2.1a): our model leverages the long-range interactions with respect to other organs (liver, stomach, spleen) and their positions to properly segment the whole pancreas region, see Fig 2.1d). Quantitative experiments conducted on two abdominal CT-

image datasets show the large performance gain brought out by U-Transformer compared to U-Net and to the local attention in [19].

2.2 Related Work

Attention Models As we said in subsection 1.3.3, only few works have been proposed and use simple attention modules [101, 102, 103, 104, 105, 19, 106]. Attention U-Net [19, 103] is one of those models and introduces an additive attention gate which aims at filtering the features coming from the skip connections, as shown in Fig. 2.3. The attention weights are computed from the gating signal coming from the previous level of the decoder and the skip connection. At the bottom, we can see the detailed attention gate and how the weights are computed. The final attention is very local because every operation is done at a pixel level.

In U-transformer, we introduce a cross-attention module in the decoder. It shares the same motivation of filtering out the skip connections based on more semantic features than in Attention U-Net. However, the attention gate shares the same limitation of local attention as the other models. On the other hand, our cross-attention is based on Transformers [29] and is able to model long-range interactions. Moreover, our MHCA is original in its design since the keys and the queries are computed from the high-level features. It differs from the standard way cross-attention is used in [29]. In our case, we are not trying to express similarities between the different U-Net levels but rather to filter the skip-connections based on the self-similarity of more semantic features. On top of that, we propose to add a Multi-Head Self-Attention (MHSA) in the bottleneck which further enforces the modeling of global interactions in our model, which are not leveraged in Attention U-Net.

Discussion on Concurrent Works. Transformer networks have not been extensively studied in medical image analysis. However, there have been several attempts in the last few months [120, 107, 121, 36, 14]. In TransUnet [107], the authors propose a method inspired by DeTr [122] integrated into a U-Net model. It could be seen as using only self-attention in the bottleneck as compared to our model which also adds cross-attention mechanisms in the skip connections. In the TransFuse model [121], the attention module is inspired by SeTr [123] where the image is first divided into patches which are then considered as tokens. Using this approach reduces considerably the input information contrary to our model which uses the complete image and could model finer global interactions. In CoTr [14], the model is based on Deformable DeTr [124] which is a very specific method aiming at reducing the memory needed by Transformers by using a "deformable" Transformer that does not compute the complete attention matrix. Instead, they use a limited number of reference points which point with an offset vector to the most important tokens but not all of them. It allows the processing of multi-scale and high-resolution features. Despite those attempts, none of them propose to use a cross-attention in

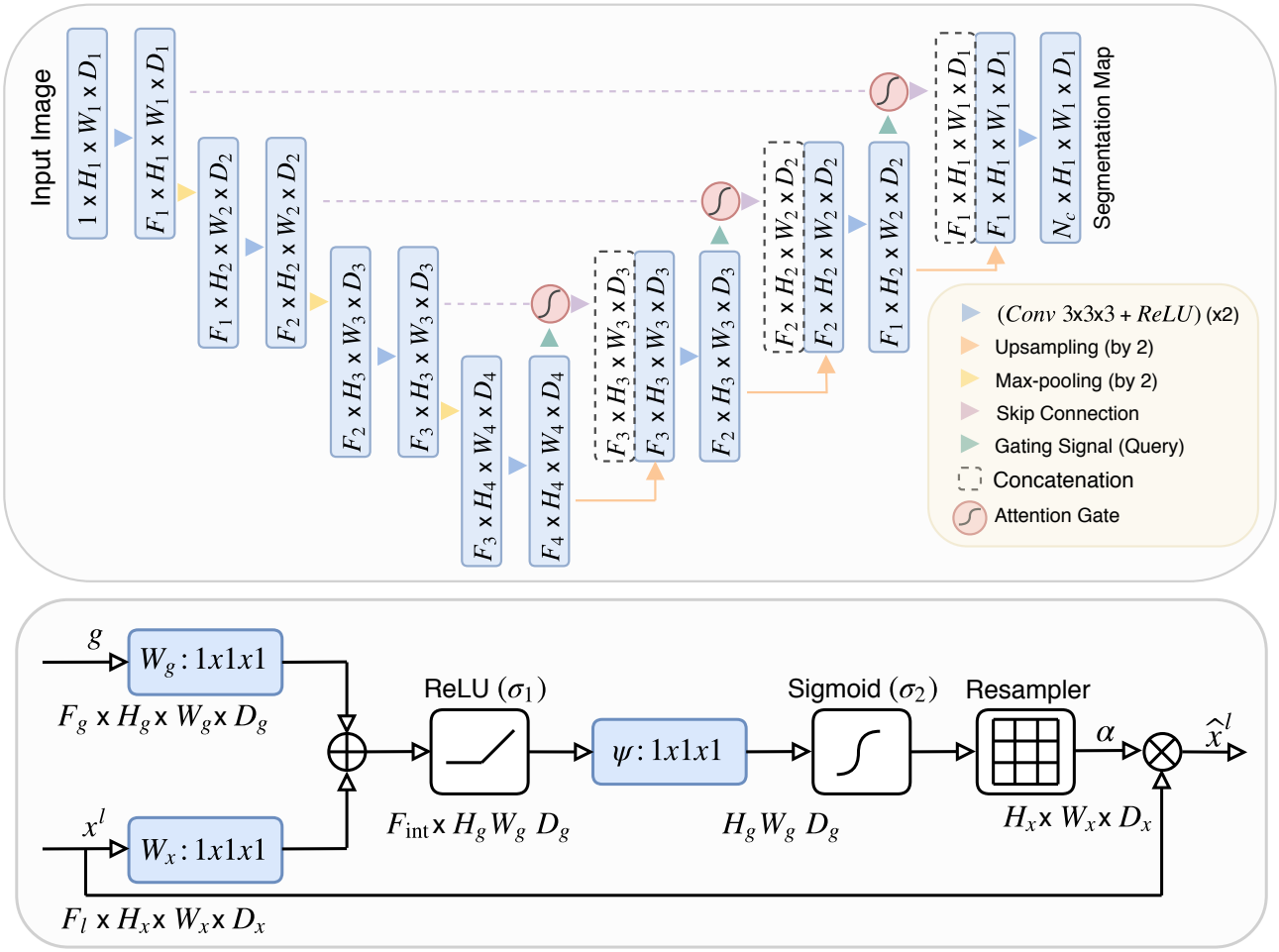


Figure 2.3: **The Attention U-Net as proposed in** [19]: The top image is the overall architecture with Attention Gates (AGs) at each skip connection. The bottom image is the attention gate mechanism with g being the gating signal (from the previous decoder block) and x the input signal (the skip connection).

a U-shape FCN to improve the spatial recovery in the decoder, contrary to U-Transformer.

2.3 The U-Transformer Network

As mentioned in Section 2.1, encoder-decoder U-shaped architectures lack global context information to handle complex medical image segmentation tasks. We introduce the U-Transformer network, which augments U-Nets with attention modules built from multi-head transformers. U-Transformer models long-range contextual interactions and spatial dependencies by using two types of attention modules (see Fig 2.4): Multi-Head Self-Attention (MHSA) and Multi-Head Cross-Attention (MHCA).

Both modules are designed to express a new representation of the input based on its self-attention in the first case (*cf.* 2.3.1) or on the attention paid to higher level features in the second (*cf.* 2.3.2).

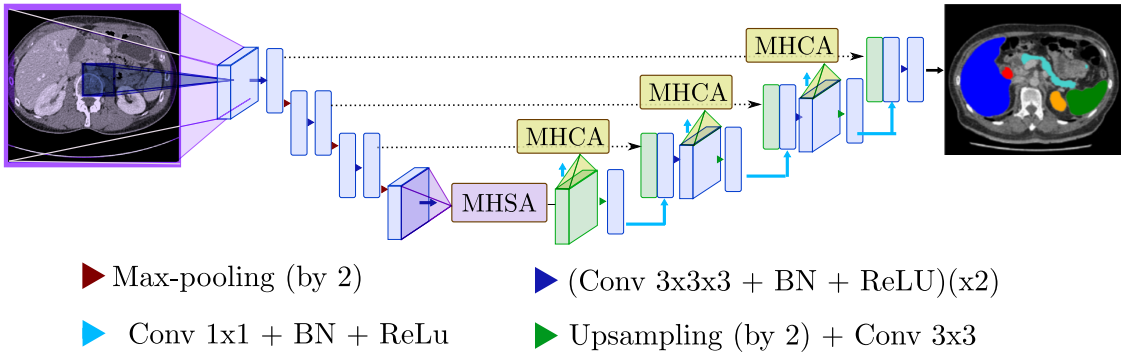


Figure 2.4: **U-Transformer** augments U-Nets with transformers to model long-range contextual interactions. The Multi-Head Self-Attention (MHSA) module at the end of the U-Net encoder gives access to a receptive field containing the whole image (shown in purple), in contrast to the limited U-Net receptive field (shown in blue). Multi-Head Cross-Attention (MHCA) modules are dedicated to combine the semantic richness in high level feature maps with the high-resolution ones coming from the skip connections.

2.3.1 Self-attention

The MHSA module is designed to extract long-range structural information from the images. To this end, it is composed of multi-head self-attention functions as described in [29] positioned at the bottom of the U-Net as shown in Figure 2.4. The main goal of MHSA is to connect every element in the highest feature map with each other, thus giving access to a receptive field including all the input images. The decision for one specific pixel can thus be influenced by any input pixel. The attention formulation is given in Equation 2.5.

2.3.2 Cross-attention

The MHSA module allows to connect every element in the input with each other. Attention may also be used to increase the U-Net decoder efficiency and in particular, enhance the lower level feature maps that are passed through the skip connections. Indeed, if these skip connections ensure to keep a high-resolution information they lack the semantic richness that can be found deeper in the network. The idea behind the MHCA module is to turn off irrelevant or noisy areas from the skip connection features and highlight regions that present a significant interest for the application. Figure 2.6 shows the cross-attention module. The MHCA block is designed as a gating operation of the skip connection

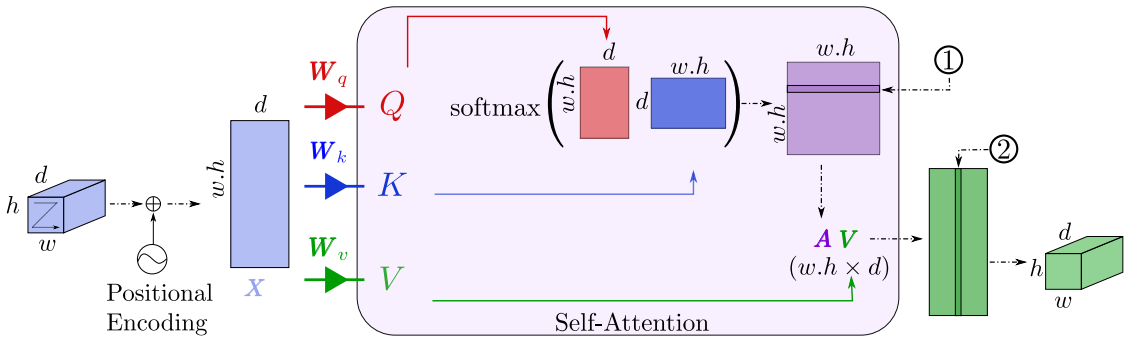


Figure 2.5: **MHSA module**: the input tensor is embedded into a matrix of queries Q, keys K and values V. The attention matrix A in purple is computed based on Q and K. (1) A line of A corresponds to the attention given to all the elements in K with respect to one element in Q. (2) A column of the value V corresponds to a feature map weighted by the attention in A.

S based on the attention given to a high level feature map Y. The computed weight values are then re-scaled between 0 and 1 through a sigmoid activation function. The resulting tensor, denoted Z in Figure 2.6, is a filter where low magnitude elements indicate noisy or irrelevant areas to be reduced. A cleaned-up version of S is then given by the Hadamard product $Z \odot S$. Finally, the result of this filtering operation is concatenated with the high level feature tensor Y. Here, the keys and queries are computed from the same source as we are designing a filtering operation whereas for NLP tasks, having homogeneous keys and values may be more meaningful. This configuration proved to be empirically more effective.

2.4 Experiments

We evaluate U-Transformer for abdominal organ segmentation on The Cancer Imaging Archive (TCIA) pancreas public dataset [30], and an Internal Multi-Organ dataset (IMO).

Accurate pancreas segmentation is particularly difficult, due to its small size, complex and variable shape, and because of the low contrast with the neighboring structures, see Fig 2.1. In addition, the multi-organ setting assesses how U-transformer can leverage attention from multi-organ annotations.

Experimental setup The TCIA pancreas dataset[1] contains 82 CT-scans with pixel-level annotations. Each CT-scan has around $181 \sim 466$ slices of 512×512 pixels and a voxel spacing of ($[0.66 \sim 0.98] \times [0.66 \sim 0.98] \times [0.5 \sim 1.0]$) mm^3.

We also experiment with an Internal Multi-Organ (IMO) dataset composed of 85 CT-scans annotated with 7 classes: liver, gallbladder, pancreas, spleen, right and left kidneys, and stomach. Each

[1] https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT

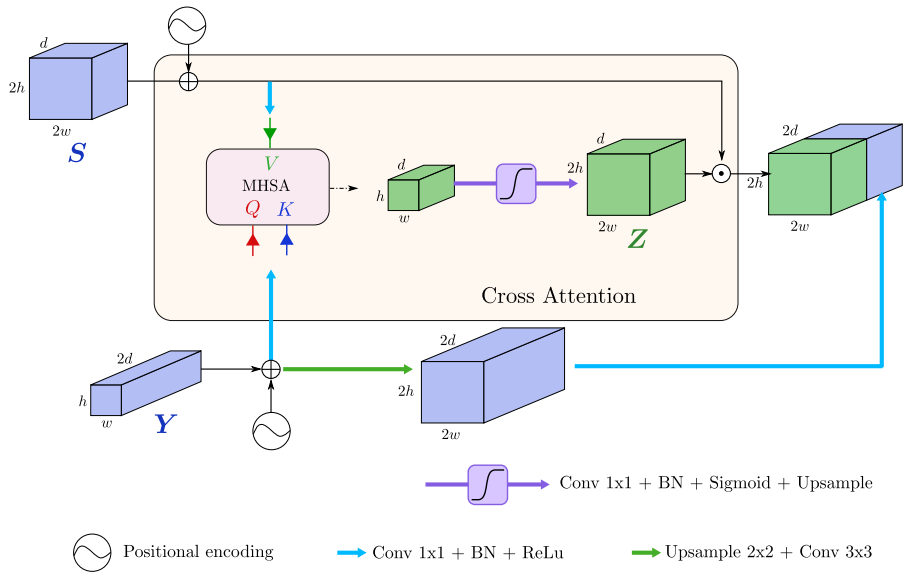


Figure 2.6: **MHCA module**: the value of the attention function corresponds to the skip connection S weighted by the attention given to the high level feature map Y. This output is transformed into a filter Z and applied to the skip connection.

CT-scan has around $57 \sim 500$ slices of 512×512 pixels and a voxel spacing of $([0.42 \sim 0.98] \times [0.42 \sim 0.98] \times [0.63 \sim 4.00])\text{mm}^3$.

All experiments follow a 5-fold cross validation, using 80% of images in training and 20% in test. We use the Tensorflow library to train the model, with Adam optimizer (10^{-4} learning rate, exponential decay scheduler).

We compare U-Transformer to the U-Net baseline [20] and Attention U-Net [19] with the same convolutional backbone for fair comparison. We also report performances with self-attention only (MHSA, section 2.3.1), and the cross-attention only (MHCA, section 2.3.2). U-Net has ~ 30M parameters, and the overhead from U-transformer is limited (MHSA ~ 5M, each MHCA block ~ 2.5M).

2.4.1 U-Transformer performances

Table 2.1 reports the performances in Dice averaged over the 5 folds and over organs for IMO. U-Transformer outperforms U-Net by 2.4pts on TCIA and 1.3pts for IMO, and Attention U-Net by 1.7pts for TCIA and 1.6pts for IMO. The gains are consistent on all folds, and paired t-tests show that the improvement is significant with $p-$values $< 3\%$ for every experiment.

Extended results. We also conducted extended experiments using nnU-Net [1] as a baseline and a

Table 2.1: Results for each method in Dice similarity coefficient (DSC, %) on TCIA and IMO. Bold indicates best performances.

Dataset	U-Net [20]	Attn U-Net [19]	MHSA	MHCA	U-Transformer
TCIA	76.13 ($\pm$ 0.94)	76.82 ($\pm$ 1.26)	77.71 ($\pm$ 1.31)	77.84 ($\pm$ 2.59)	**78.50** ($\pm$ 1.92)
IMO	86.78 ($\pm$ 1.72)	86.45 ($\pm$ 1.69)	87.29 ($\pm$ 1.34)	87.38 ($\pm$ 1.53)	**88.08** ($\pm$ 1.37)

training pipeline. nnU-Net is a stronger and deeper version of UNet, well optimized and trained in a tailored training pipeline for medical image segmentation. It reaches the best results on multiple tasks. We evaluated our approach using the nnU-Net authors' Github code on TCIA using 3 fold and following the experimental setup in [1]. As presented in Tab. 2.2 our results show a gain of 1pt (84.08 vs 83.09 in Dice), which is a large improvement given the strong baseline, and statistically significant with a paired t-test (p=0.023). This highlights that our MHSA/MHCA modules improve performances over state-of-the-art convolutional models.

Table 2.2: Extended results in Dice (%) on TCIA using nnU-Net [1] baseline and experimental setup. In this setup, U-Transformer has more layers, and the MHSA is applied also on more layers. The number of layers match nnU-Net architecture.

Dataset	nnU-Net [1]	MHSA	MHCA	U-Transformer
TCIA	83.09 ($\pm$ 1.23)	83.78 ($\pm$ 1.12)	83.41 ($\pm$ 1.08)	84.08 ($\pm$ 1.17)

Figure 2.7 provides a qualitative segmentation comparison between U-Net, Attention U-Net and U-Transformer. We observe that U-Transformer performs better on difficult cases, where the local structures are ambiguous. For example, in the second row, the pancreas has a complex shape that is missed by U-Net and Attention U-Net but U-Transformer successfully segments the organ.

In Table 2.1, we can see that the self-attention (MHSA) and cross-attention (MHCA) alone already outperform U-Net and Attention U-Net on TCIA and IMO. Since MHCA and Attention U-Net apply attention mechanisms at the skip connection level, it highlights the superiority of modeling global interactions between anatomical structures and positional information instead of the simple local attention in [19]. Finally, the combination of MHSA and MHCA in U-Transformer shows that the two attention mechanisms are complementary and can collaborate to provide better segmentation predictions.

Table 2.3 details the results for each organ on the multi-organ IMO dataset. This further highlights the interest of U-Transformer, which significantly outperforms U-Net and Attention U-Net for the most challenging organs: pancreas: +3.4pts, gallbladder: +1.3pts and stomach: +2.2pts. This validates the capacity of U-Transformer to leverage multi-label annotations to drive the interactions between anatomical structures and use easy organ predictions to improve the detection and delineation of more

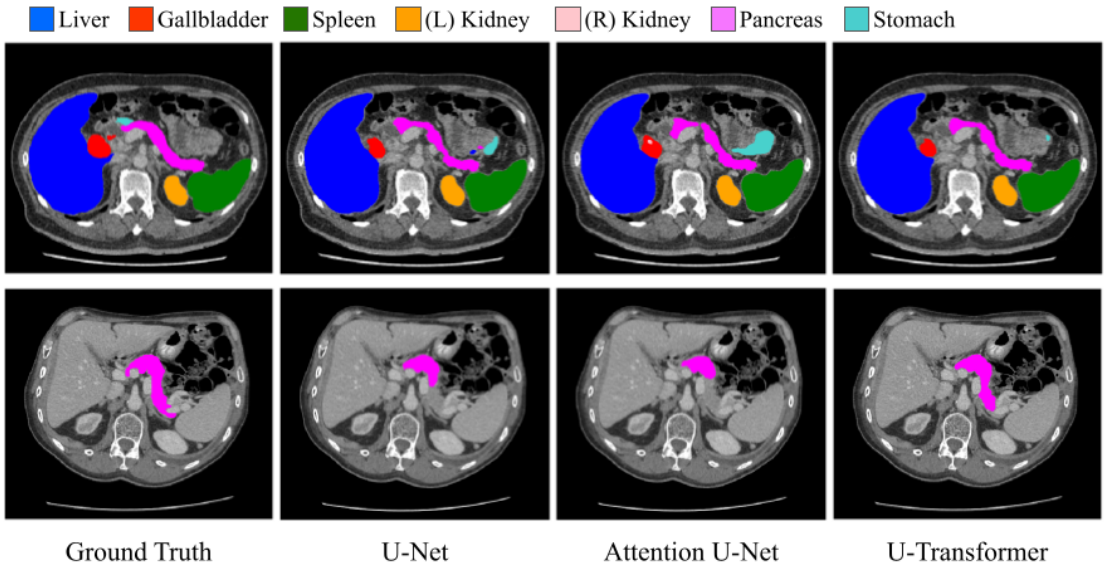


Figure 2.7: Segmentation results for U-Net [20], Attention U-Net [19] and U-Transformer on the multi-organ IMO dataset (first row) and on TCIA pancreas (second row).

difficult ones. We can note that U-Transformer is better for every organ, even the liver which has a high score $> 95\%$ with U-Net.

Table 2.3: Results on IMO in Dice similarity coefficient (DSC, %) detailed per organ.

Organ	U-Net [20]	Attn U-Net [19]	MHSA	MHCA	U-Transformer
Pancreas	69.71 ($\pm$ 3.74)	68.65 ($\pm$ 2.95)	71.64 ($\pm$ 3.01)	71.87 ($\pm$ 2.97)	**73.10** ($\pm$ 2.91)
Gallbladder	76.98 ($\pm$ 6.60)	76.14 ($\pm$ 6.98)	76.48 ($\pm$ 6.12)	77.36 ($\pm$ 6.22)	**78.32** ($\pm$ 6.12)
Stomach	83.51 ($\pm$ 4.49)	82.73 ($\pm$ 4.62)	84.83 ($\pm$ 3.79)	84.42 ($\pm$ 4.35)	**85.73** ($\pm$ 3.99)
Kidney(R)	92.36 ($\pm$ 0.45)	92.88 ($\pm$ 1.79)	92.91 ($\pm$ 1.84)	92.98 ($\pm$ 1.70)	**93.32** ($\pm$ 1.74)
Kidney(L)	93.06 ($\pm$ 1.68)	92.89 ($\pm$ 0.64)	92.95 ($\pm$ 1.30)	92.82 ($\pm$ 1.06)	**93.31** ($\pm$ 1.08)
Spleen	95.43 ($\pm$ 1.76)	95.46 ($\pm$ 1.95)	95.43 ($\pm$ 2.16)	95.41 ($\pm$ 2.21)	**95.74** ($\pm$ 2.07)
Liver	96.40 ($\pm$ 0.72)	96.41 ($\pm$ 0.52)	96.82 ($\pm$ 0.34)	96.79 ($\pm$ 0.29)	**97.03** ($\pm$ 0.31)

2.4.2 U-Transformer analysis and properties

Positional encoding and multi-level MHCA. The Positional Encoding (PE) allows to leverage the absolute position of the objects in the image. Table 2.4 shows an analysis of its impact, on one fold on both datasets. For MHSA, the PE improves the results by +0.7pt for TCIA and +0.6pt for IMO. For MHCA, we evaluate a single level of attention with and without PE. We can observe an improvement of +1.7pts for TCIA and +0.6pt for IMO between the two versions.

Table 2.4 also shows the favorable impact of using multi *vs* single-level attention for MHCA: +1.8pts for TCIA and +0.6pt for IMO. It is worth noting that Attention U-Net uses multi-level attention but remains below MHCA with a single level. Figure 2.8 shows attention maps at each level of U-Transformer: level 3 corresponds to high-resolution features maps, and tends to focus on more specific regions compared to the first levels.

Table 2.4: Ablation study on the positional encoding and multi-level on one fold of TCIA and IMO.

			MHSA		MHCA		
	U-Net	Attn U-Net	wo PE –	w PE	1 lvl wo PE –	1 lvl w PE –	multi-lvl w PE
TCIA	76.35	77.23	78.17	**78.90**	77.18	78.88	**80.65**
IMO	88.18	87.52	88.16	**88.76**	87.96	88.52	**89.13**

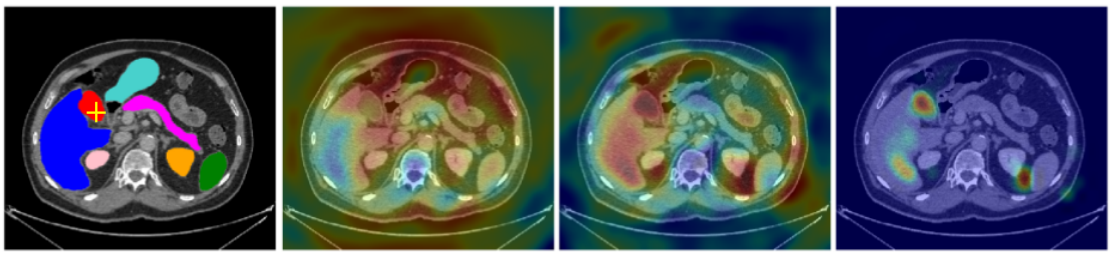


Figure 2.8: Cross-attention maps for the yellow-crossed pixel (left image).

Further analysis. To further analyze the behavior of U-Transformer, we evaluate the impact of the number of attention heads for MHSA Fig. 2.9: more heads lead to better performances, but the biggest gain comes from the first head (i.e. U-Net to MHSA). Finally, the evaluation of U-Transformer with respect to the Hausdorff distance Tab. 2.5 follows the same trend as with the Dice score. This highlights the capacity of U-Transformer to reduce prediction artifacts by means of self- and cross-attention. In addition, we have evaluated our method on a TCIA multiorgan extension which gives the same trends as with our IMO Table 5.3.

Table 2.5: Hausdorff Distances (HD) for the different models

Dataset	U-Net	Attn U-Net	U-Transformer
TCIA	13.61 ($\pm$ 2.01)	12.48 ($\pm$ 1.36)	**12.34** ($\pm$ 1.51)
IMO	12.06 ($\pm$ 1.65)	12.13 ($\pm$ 1.58)	**12.00** ($\pm$ 1.32)

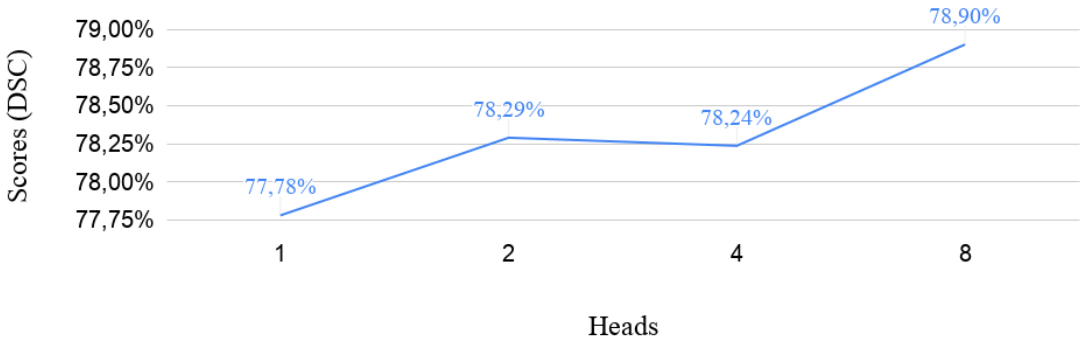


Figure 2.9: Evolution of the Dice Score on TCIA (fold 1) when the number of heads varies between 0 and 8 in MHSA.

2.5 Conclusion

This chapter introduces the U-Transformer network, which augments a U-shaped FCN with Transformers. We propose to use self and cross-attention modules to model long-range interactions and spatial dependencies. We highlight the relevance of the approach for abdominal organ segmentation, especially for small and complex organs. To enhance the ability of U-Net to capture more context, the utilization of 3D images would be appropriate. U-transformer is limited to 2D segmentation and the self-attention module is in the bottleneck modeling only coarse interactions. The extension of this method to 3D images while keeping global interactions at each level is an open question and the core of the next chapter.

Chapter 3

Full Contextual Attention for Multi-resolution Transformers in Semantic Segmentation

Contents

3.1 Introduction **75**
3.2 Related work **76**
3.3 The GLAM Method **77**
3.3.1 Global attention multi-resolution transformers 78
3.4 Experiments **82**
3.4.1 Experimental Settings 82
3.4.2 GLAM performance 83
3.4.3 Additional results 85
3.4.4 Model Analysis 86
3.4.5 Visualizations. 89
3.5 Conclusion **91**

Chapter summary

The quadratic complexity of self and cross-attention in U-Transformer renders it ineffective in handling the high dimensionality of 3D medical images. This issue also impacts the architecture of the model as the self-attention can only be performed in the bottleneck and the cross-attention requires a downsampling pre-processing. Thus, U-Transformer can't model long-range interaction with high-resolution features which limits its efficiency. This chapter aims to address these limitations

by incorporating long-range interaction modeling even in high-resolution feature maps. We present GLAM (GLobal Attention Multi-resolution transformers) as a solution to address the aforementioned limitations. The core idea of GLAM is to incorporate global tokens in the model, influenced by the class token[93, 125], that serves as pivotal elements for the transmission of information across various regions of the image at each stage. GLAM includes learnable global tokens, which unlike classical transformer-based methods can model interactions between all image regions, and extract powerful representations during training. GLAM is a generic module that can be integrated into most existing windowed Transformers backbones. Consequently, this integration of GLAM engenders extensive interactions over long distances, which were previously absent in the model. Extensive experiments show that GLAM-Swin or GLAM-Swin-UNet exhibit substantially better performances than their vanilla counterparts on ADE20K and Cityscapes. Moreover, GLAM can be used to segment large 3D medical images, and GLAM-nnFormer achieves new state-of-the-art performance on the BCV dataset.

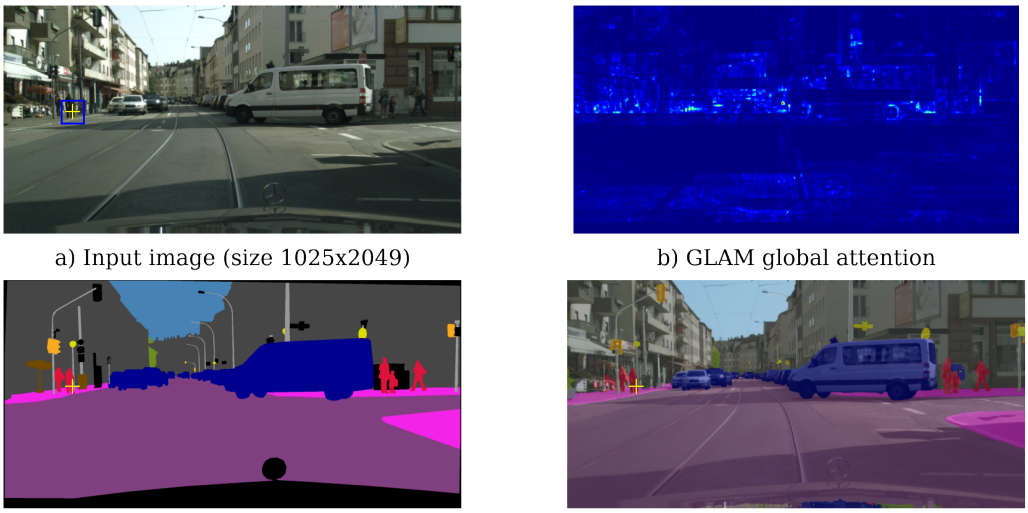
3.1 Introduction

The main appeal of transformers is their ability to grasp long-range interactions, which is a crucial point for semantic segmentation. However, this strategy is not easily scalable to high-resolution images involving a large number of patches, due to the quadratic complexity of the transformer's attention module. For instance, The U-Transformer presented in Chapter 2 only applies self-attention in the bottleneck, and cross-attention uses a downsampling operation which degrades the spatial information. A simple and efficient strategy to tackle this limitation is to rely on multi-resolution approaches, where the attention in high-resolution feature maps is computed on sub-windows. There have been various recent attempts in this direction [17, 114, 126, 112, 21]. However, they limit the interactions of high-resolution features to within each window. These limitations will be detailed in the related work section.

We introduce an approach for semantic segmentation that incorporates global attention in multi-resolution transformers (GLAM). The GLAM module enables full-range interactions to be modeled at all scales of a multi-resolution transformer. As illustrated in Fig. 3.1, incorporating GLAM into the Swin architecture [17] enables to jointly capture fine-grained spatial information in high-resolution feature maps and global context, where both elements are crucial for proper segmentation in complex scenes. This concept is illustrated in Fig. 3.1 where Fig. 3.1a) shows an input image, and Fig. 3.1b) shows the self-attention map provided by GLAM in the highest-resolution feature map for the pedestrian region pointed out by the yellow cross in Fig. 3.1a). We can see that the attention map involves long-range interactions between other visual structures (cars, buildings), in contrast to the Swin baseline, where the window attention at a high-resolution feature map is limited to the small rectangular region in Fig. 3.1a). Consequently, GLAM has exploited longer-range interactions to successfully segment the image, as shown in 3.1d).

To achieve this goal, we have made the following novel contributions:

- We introduce the GLAM transformer, able to represent full-range interactions between all local features at all resolution levels. The GLAM transformer is based on learnable global tokens interacting between all visual features. To fully take into account the global context, we also design a non-local upsampling scheme (NLU) which is inspired by the cross-attention used in Chapter 2 but extend the idea by incorporating a full transformer module.
- GLAM is a generic module that can be incorporated into any multi-resolution transformer. It consists of a succession of two transformers applied on the merged sequence of global and visual tokens and in-between global tokens. We highlight that the GLAM transformer can represent full-range interactions between image regions at all scales while retaining memory



a) Input image (size 1025x2049) b) GLAM global attention

c) Ground Truth d) GLAM prediction

Figure 3.1: When segmenting the high-resolution image in a) with state-of-the-art multi-resolution transformers, *e.g.* Swin [17], the attention in the highest-resolution feature maps is limited to a small spatial region, *i.e.* the blue square for the yellow-crossed pedestrian. Our method incorporates GLobal Attention in Multi-resolution transformers (GLAM). The GLAM attention map for the pedestrian in a) is depicted in b): it captures both fine-grained spatial information and long-range interactions, enabling successful segmentation, as shown in d).

and computational efficiency. Beyond spatial interactions, global tokens also model the expected scene composition.

- Experiments on various generic (ADE20K) [127], autonomous driving (Cityscape) [32] and medical (BCV) datasets [128] show the important and systematic gain brought by GLAM when included into existing state-of-the-art multi-resolution transformers including Swin, Swin-UNet, and nn-Former. We also show that GLAM outperforms state-of-the-art methods on BCV. Finally, ablation studies, model analysis, and visualizations are presented to assess the behavior of GLAM.

3.2 Related work

Several recent approaches proposed adaptations of the vanilla ViT architecture. In particular, some architectures rely on multi-resolution processing. T2T ViT [129] constructs richer semantic feature map through token aggregation while TnT [130] and crossViT [131] uses two transformers for fine and coarse resolution. PvT [114] is the first backbone with a fully pyramidal architecture that is based on windowed transformers, allowing to process the images at fine resolution and to build rich

feature maps while reducing the spatial complexity. Other methods kept this hierarchical approach while improving information sharing between the windows. Swin [17] and its variant [21, 2] proposed to use shifted windows, Twins [132] uses interleaved fine and coarse resolution transformers, and CvT [133] replaces linear embedding with convolutions.

These methods are based on the fact that self-attention cannot be applied to long sequences *e.g.* patches from high-resolution images because the computation of the attention matrix has quadratic memory complexity. To allow high-resolution processing and thus long sequences of small patches, windowed transformers treat the image as a batch of non-overlapping windows [17, 114, 112, 126]. This approach is combined with a pooling strategy [21, 17, 114, 126] and is well suited to build a multi-resolution encoder, able to produce rich semantic maps. Multi-resolution backbones are built by chaining windowed transformer blocks and downsampling. These hierarchical architectures manage to build larger receptive fields in deeper layers, similar to CNNs. This, however, does not guarantee a global receptive field and the maximal receptive field depends on the model's depth. More importantly, this process introduces a major modification to the transformer modules. At a finer resolution, only local interactions are considered. With this modification, the processing of isolated patches by self-attention may not be as effective as global self-attention performed on the full image.

3.3 The GLAM Method

The main idea in GLAM is to provide a way to represent full range interactions at all feature map resolutions, which is impossible in vanilla models, especially in high-resolution feature maps, due to the quadratic complexity of attention transformers.

GLAM is illustrated in Fig. 3.2, where it has been added to the Swin-UNet architecture [17]. Note that GLAM can be included in various multi-resolution architectures, *e.g.* Swin [21] or PvT [114] and is also applicable for 3D segmentation, *e.g.* nn-Former [2]. The core idea in GLAM is to design global tokens (in red in Fig. 3.2), which are leveraged into a succession of two attention steps: first, between visual tokens in each window independently and, second, between global tokens among different windows. We show in Sec. 3.3.1 that this design enables to represent full range interactions between all image regions at all scales, and also external information useful for segmentation while retaining efficiency. We also introduce a non-local upsampling scheme (NLU) to extend the full context modeling in U-shape architectures and to provide an efficient interpolation of rich semantic feature maps in an associated decoder.

As shown in Fig. 3.2, GLAM can be included into any multi-resolution transformer architecture [17, 114, 126, 112, 21, 2].

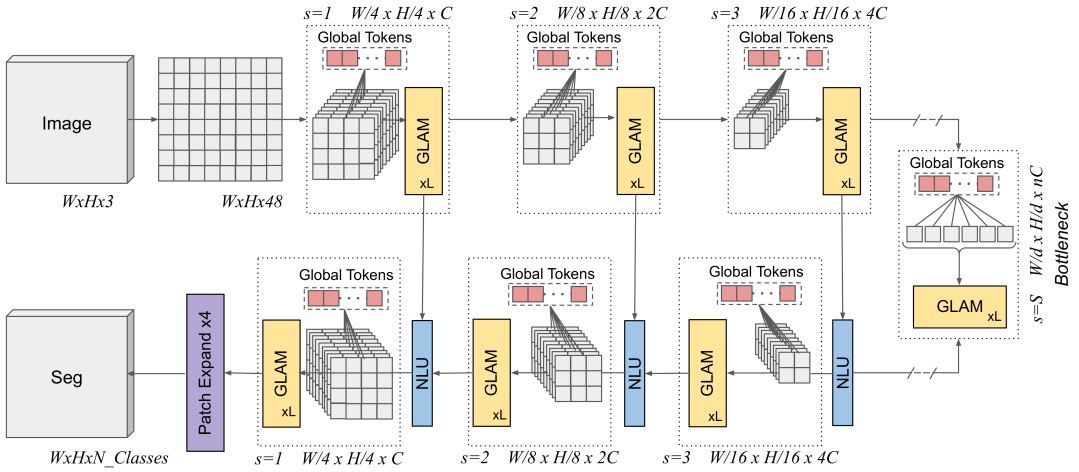


Figure 3.2: The GLAM module for modeling full-range interaction in multi-resolution transformers. GLAM is included at each resolution level of any multi-resolution transformer architecture, *e.g.* Swin-UNet [17] or Swin-UperNet [17]. GLAM includes learnable global tokens, which are leveraged into a succession of two attention steps. We show that this design can indirectly represent long-range interactions between all image regions at all scales, and also external information useful for segmentation while retaining efficiency. We also introduce a non-local upsampling scheme (NLU) to extend the global context modeling in full transformer U-shape architectures such as [21, 2].

3.3.1 Global attention multi-resolution transformers

We show how the GLAM module can provide global attention in all feature maps of multi-resolution transformers. The GLAM transformer is illustrated in Fig. 3.3, consisting of a sequence of L transformer blocks, processing visual tokens in each region of the multi-resolution maps (shown in blue in Fig. 3.3) and global tokens (shown in red in Fig. 3.3).

The basic idea behind GLAM is to associate global tokens to each window that is responsible to encapsulate the local information and transmit it to other image regions by computing MSA between all global tokens. Thus, when information is processed at the window scale, the visual tokens embedding incorporates useful long-range information.

Global Tokens. Global tokens lie at the core of Global Attention (GA). They are specific tokens concatenated to each window and are responsible for communication between windows. We define as N_w the number of windows in the feature map, N_p as the number of patches per window, and $\{\boldsymbol{v}_k^l\}_{1\leq k\leq N_w}$ as the sequence of windows after being processed by the l^{th} GLAM-transformer block. We define as $\{\boldsymbol{g}_k^l\}_{1\leq k\leq N_w}$ the sequence of N_g-dimensional global tokens associated to each window. The initialization of the global tokens $\{\boldsymbol{g}_k^0\}_{1\leq k\leq N_w}$ is the same for all windows and is learned by the model. The input of the l^{th} transformer block, defined as $\boldsymbol{z}^l$, is a batch of tokens from each

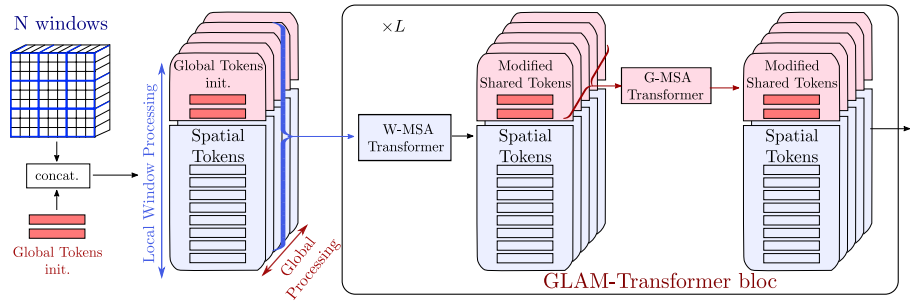


Figure 3.3: **GLAM-Transformer:** as in multi-resolution approaches, each input feature map is divided into N_w non overlapping windows (blue). The core idea in GLAM is to design learnable global tokens (in red). The visual tokens from each window are concatenated with the global tokens and processed through a local window transformer (W-MSA). Every W-MSA is followed by a global transformer (G-MSA), where global tokens between different windows interact with each other, which brings a global representation to each window. These two steps give the GLAM-Transformer block; Multiple blocks are chained at every hierarchy level in typical multi-resolution transformer backbones. We show that global tokens learned from GLAM-Transformer indirectly model global interactions between all visual tokens in all widows. The global tokens are also able to represent extra learnable knowledge beyond the patch interactions in a single image.

window concatenated with the corresponding global tokens, *i.e.* $\boldsymbol{z}^l \in \mathbb{R}^{N_w \times (N_g+N_p) \times C}$ with C being the dimension of the tokens. Consequently, the elements in the batch have the form:

$$\forall k \in [1..N_w], \boldsymbol{z}_k^l = \begin{bmatrix} \boldsymbol{g}_k^l \\ \boldsymbol{v}_k^l \end{bmatrix} \in \mathbb{R}^{(N_g+N_p) \times C}. \tag{3.1}$$

GLAM-Transformer. The communication between windows at a given hierarchy level is obtained through the interaction of global tokens. At each block l of the GLAM-transformer, there are two steps: *i)* visual tokens grasp their local statistics through a local window transformer (W-MSA), and *ii)* the global tokens are re-embedded by a global transformer (G-MSA), where global tokens from different windows interact with each other. Formally, the l^{th} GLAM-transformer block inputs $\boldsymbol{z}^{l-1}$ and outputs $\boldsymbol{z}^l$ by the succession of a W-MSA and a G-MSA step:

$$\begin{aligned} \hat{\boldsymbol{z}}^l &= \text{W-MSA}(\boldsymbol{z}^{l-1}), \\ \boldsymbol{g}^l &= \text{G-MSA}(\hat{\boldsymbol{g}}^l), \\ \boldsymbol{z}^l &= \left[\boldsymbol{g}_k^{l^T} \hat{\boldsymbol{v}}_k^{l^T} \right]^T \end{aligned} \tag{3.2}$$

We define as $\boldsymbol{A}_r^l$ the attention matrix for the window r in the transformer block l. We introduce the following decomposition to express the attention with respect to the global and local tokens:

$$\boldsymbol{A}_r^l = \begin{bmatrix} \boldsymbol{A}_{r,gg}^l & \boldsymbol{A}_{r,gv}^l \\ \boldsymbol{A}_{r,vg}^l & \boldsymbol{A}_{r,vv}^l \end{bmatrix}. \tag{3.3}$$

The square matrices $\boldsymbol{A}^l_{r,gg} \in \mathbb{R}^{N_g \times N_g}$ and $\boldsymbol{A}^l_{r,vv} \in \mathbb{R}^{N_p \times N_p}$ give the attention from the global token and the spatial tokens on themselves respectively. The matrices $\boldsymbol{A}^l_{r,gv} \in \mathbb{R}^{N_g \times N_p}$ and $\boldsymbol{A}^l_{r,vg} \in \mathbb{R}^{N_p \times N_g}$ are the cross attention matrices between local and global tokens. We define as $\boldsymbol{B}^l \in \mathbb{R}^{(N_w \cdot N_g) \times (N_w \cdot N_g)}$ the global attention matrix from all the global token sequence and $\boldsymbol{B}^l_{ij} \in \mathbb{R}^{N_g \times N_g}$ as the sub-matrices giving the attention between the global tokens of windows i and j.

GLAM-Transformer properties. Putting aside the value matrix, the W-MSA gives the following embedding $\hat{\boldsymbol{g}}^l_r$ from $\boldsymbol{g}^{l-1}_r$:

$$\hat{\boldsymbol{g}}_r = \boldsymbol{A}^l_{r,gg}\boldsymbol{g}^{l-1}_r + \boldsymbol{A}^l_{r,gv}\boldsymbol{v}^{l-1}_r \,. \tag{3.4}$$

The G-MSA, *i.e.* the MSA on the sequence of global tokens gives the following embeddings:

$$\begin{aligned} \boldsymbol{g}^l_r &= \sum_{n=1}^{N_w} \boldsymbol{B}^l_{rn}\hat{\boldsymbol{g}}^l_n \\ &= \sum_{n=1}^{N_w} \boldsymbol{B}^l_{rn}(\boldsymbol{A}^l_{r,gg}\boldsymbol{g}^{l-1}_r + \boldsymbol{A}^l_{r,gv}\boldsymbol{v}^{l-1}_r) \,. \end{aligned} \tag{3.5}$$

From Eq. 3.5, we have the expression of the global token for a window r processed by the l G-MSA block transformer. Developing this formulation we obtain the following expression for the k^{th} global token in the r^{th} window:

$$\begin{aligned} g^l_{k,r} &= \sum_{r'=1}^{N_w}\sum_{j=1}^{N_g} b_{k,r,j,r'}\left(\sum_{i=1}^{N_g+N_p} a_{j,r',i} z^{l-1}_{i,r'}\right) \\ &= \sum_{r'=1}^{N_w}\sum_{j=1}^{N_g} b_{k,r,j,r'}\left(\sum_{i=1}^{N_g} a_{j,r',i} g^{l-1}_{i,r'} \right. \\ &\qquad \left. + \sum_{i=1}^{N_p} a_{j,r',i+N_g} v^{l-1}_{i,r'}\right). \end{aligned} \tag{3.6}$$

The variables $z_{i,r}$, $g_{i,r}$ and $v_{i,r}$ corresponds respectively to the visual, global or generic token i in window r. $a_{j,r,i}$ is the attention coefficient given by the token j to the token i inside the window r. $b_{j,r,i,r'}$ is the attention coefficient from the global token j in the window r to the global token j in the window r'. Re-arranging the indices of equation 3.6 leads to the following expression for the k^{th} global token in the r^{th} window:

$$g^l_{k,r} = \sum_{r'=1}^{N_w} \sum_{i=1}^{N_p} \left(\sum_{j=1}^{N_g} b_{k,r,j,r'}\, a_{j,r',(i+N_g)}\, v^{l-1}_{i,r'} \right) + \sum_{r'=1}^{N_w} \left(\sum_{j=1}^{N_g} b_{k,r,j,r'} \sum_{i=1}^{N_g} a_{j,r',i}\, g^{l-1}_{i,r'} \right) \tag{3.7}$$

This leads to a global attention matrix $\boldsymbol{G}_k \in \mathbb{R}^{(N_w \cdot N_p)\times(N_w \cdot N_p)}$ associated to the k^{th} global token given by $[\boldsymbol{G}_k]_{r',i} = \sum_{j=1}^{N_g} b_{k,r,j,r'} a_{j,r',(i+N_g)} + \sum_{j=1}^{N_g} b_{k,r,j,r'} \sum_{i=1}^{N_g} a_{j,r',i}$. Eq.3.7 gives the embedding of the global token $g^l_{k,r}$ at the l^{st} GLAM-transformer block, with respect to all visual tokens in all feature map windows $v^{l-1}_{i,r'}$ (first row), and all global tokens $g^{l-1}_{i,r'}$ (second row). This rewriting shows that the global embedding $g^l_{k,r}$ captures interactions between all image regions independently of the resolution. The different terms in the decomposition are interpreted as an attention map associated with each image region. This is the visualization shown in Fig.3.1: the row of the first term corresponds to patch-based attention which depends on all the tokens of the feature map, while the second row represents window-based attention.

Overall, global tokens embedded with GLAM-transformers provide a way for information propagation across all windows (first row in Eq.. 3.7), but also global information (second row) that goes beyond matching visual features in a single image. Especially, this represents global and learned information across the dataset and can be leveraged as a stabilizing effect in SA, because the information is shared not only from the input but from all the windows in the dataset. This makes them a powerful tool to interpret isolated tokens and to take advantage of redundant structures in the data.

GLAM-Transformer complexity. The computational complexity of an MSA module for an image I divided into $h \times w$ patches has quadratic scaling with respect to the image area hw. The windowed approach W-MSA only depends on $N_p hw$. The complexity of both methods is given by:

$$\Omega(\text{MSA}(I)) = 4hwc^2 + 2(hw)^2 c \tag{3.8}$$

$$\Omega(\text{W-MSA}(I)) = 4hwc^2 + 2N_p hwc \tag{3.9}$$

This makes the W-MSA scalable to a large number of patches where the MSA can not be computed. With few global tokens, the global attention adds only a few numbers of operations as it corresponds to adding N_g tokens in each window and performing MSA over a sequence of length $N_g \times N_w$. It is also worth noting that the global tokens add a limited memory overhead as they do not require any more activation saving and only add a few elements in the attention matrix from each transformer block.

Non-Local Upsampling. We introduce a Non-Local Upsampling (NLU) module for a full transformer

decoder such as [21, 2]. NLU is designed to upsample the semantic features based on all the tokens coming from the skip connection, by drawing inspiration for non-local means [134].

The proposed NLU is illustrated in the supplementary material. To perform the upsampling, the skip connections are embedded into a query matrix of size $(4N_p) \times C$ while the semantic low-resolution features are embedded into the keys and values of size $N_p \times C$. The projection of the values on the resulting attention matrix has the size $(4N_p) \times C$.

3.4 Experiments

3.4.1 Experimental Settings

Datasets. We evaluated our method on three different semantic segmentation datasets: ADE20K [31], Cityscapes [32] and BCV [33]. ADE20K is a scene parsing dataset composed of 20,210 images with 150 object classes. Cityscapes contains driving scenes and is composed of 5,000 images annotated with 19 different classes. BCV is an abdominal organ segmentation dataset that includes 30 CT scans which are 3D volumes annotated with 8 abdominal organs.

Implementation details. GLAM models were implemented into the mmseg [135] codebase and the models were trained on 8 Tesla V100 GPUs. The layers were pretrained on ImageNet-1K and standard augmentation was used: random crop, rotations, translations, *etc.* We used the Adam optimizer with a weight decay of 0.01 and a polynomial learning rate scheduler starting from 0.00006 and with a factor of 1.0. The reported segmentation performances are mean Intersection over Union (mIoU) for ADE20k and Cityscapes and Dice Similarity Score (DSC) for BCV.

Training details: ADE20K and Cityscapes. For both ADE20K and Cityscapes, we implemented GLAM into the mmseg codebase [135]. All experiments ran on 8 Tesla V100 GPUs with 32GB and a batch size of 16 using data augmentation from the mmseg framework: random horizontal flipping, random re-scaling within ratio range [0.5, 2.0] and random photometric distortion. GLAM is implemented into the Swin and Swin-UNet models. Therefore, we were able to use the pre-trained weights from the respective models on ImageNet-1k [136]. For the case of the Swin-UNet backbone, we keep the same strategy as in [21] and duplicate symmetrically the encoder's weights to the decoder before fine-tuning. The added NLU and G-MSA modules could not benefit from this strong pre-training and their parameters were initialized randomly. Thanks to their integration into the overall architecture and the limited parameter increase they represent, this did not impact the good performances of the GLAM models. Complete pre-training on ImageNet of the GLAM backbones may however lead to even higher scores. The chosen optimizer is Adam with a weight decay of 0.01 and a polynomial learning rate scheduler starting from 0.00006 and with a factor of 1.0. The images

in train are cropped at a size of 512×512 for ADE20K and 768×768 for Cityscapes. In validation the complete image is provided.

Training details: BCV. BCV is a medical image dataset composed of abdominal CT-scans. Thus, the models aren't pretrained on ImageNet as for ADE20K or Cityscapes. However, we integrated our experiments in the nnUNet framework that integrates an efficient training procedure. We followed the nnFormer model and used the SGD optimizer with an initial learning rate of 0.01. We employ a polynomial learning rate scheduler and a weight decay of 3e-5. The loss function is a combination of the cross entropy and dice. Similarly to nnFormer, the numbers of heads used in the encoder stages are [6, 12, 24, 48]. The training is performed through 1000 epochs where each image is cropped at a size of $(128 \times 128 \times 64)$, as it is classically done for semantic segmentation over large 3D medical images, and in validation, we use a sliding window on the complete input volume.

3.4.2 GLAM performance

Table 3.1: **GLAM Improvements on various multi-resolution transformers.** Performances are evaluated with respect to mIoU for ADE20k and Cityscapes and average DSC for BCV.

Dataset	Method	Size	Score
ADE20K	Swin-UNet [21]	Tiny	42.75
	GLAM-Swin-UNet	Tiny	**44.19**
	Swin-UNet [21]	Small	47.49
	GLAM-Swin-UNet	Small	**47.90**
	Swin-UNet [21]	Base	47.85
	GLAM-Swin-UNet	Base	**49.10**
	Swin-UperNet[17]	Tiny	43.69
	GLAM-Swin-UperNet	Tiny	**44.16**
	Swin-UperNet [17]	Small	47.72
	GLAM-Swin-UperNet	Small	**47.75**
	Swin-UperNet [17]	Base	47.99
	GLAM-Swin-UperNet	Base	**48.44**
Cityscapes	Swin-UperNet [17]	Tiny	78.24
	GLAM-Swin-UperNet	Tiny	**78.64**
	Swin-UperNet [17]	Base	80.79
	GLAM-Swin-UperNet	Base	**81.47**
	Swin-UNet [21]	Tiny	77.43
	GLAM-Swin-UNet	Tiny	**78.29**
BCV	nnFormer [2]	Tiny	87.40
	GLAM-nnFormer	Tiny	**88.60**

GLAM in multi-resolution transformers. GLAM is well suited to work with window transformers such as PvT [114, 112] or Swin [17] as well as its variants [21, 2]. Due to the top performances of Swin, we incorporated GLAM into this backbone to compute the segmentation of 2D datasets leading to two models: GLAM-Swin-UperNet and GLAM-Swin-UNet. The first one is a hybrid model combining a transformer backbone and a CNN head [21, 80] while the second one is a full transformer model with a decoder symmetric to the encoder [21]. For 3D images, GLAM was plugged into nnFormer [2] which is designed similarly to Swin-UNet for 3D medical image segmentation. The performances of the Swin and GLAM models are presented in Table 3.1. GLAM models exhibit important and consistent performance gains compared to their vanilla counterparts, either on small or larger models: *e.g.* $\sim$ +1.5pt gain on ADE20K with Swin-UNet (Base or Tiny), and +1.2pt on BCV on the recent nn-Former model.

State-of-the-art comparison. We now compare the GLAM-Swin models with existing approaches on BCV [33], ADE20K [31] and Cityscapes [32].

BCV. Table 3.2 reports our results and recent baselines for 3D medical segmentation. GLAM-nnFormer significantly outperforms all other existing methods by at least 1.2% average Dice. To the best of our knowledge, GLAM-nnFormer outperforms state-of-the-art on the BCV dataset.

ADE20K and Cityscapes. Table 3.3 summarizes our results. To be fair, we compared models up to $\sim$ 150M parameters, and we report the top performances from the mmseg [135] benchmark for all methods, with 160K training epochs for all methods. Moreover, we compared only methods trained on 768×768 resolution images on Cityscapes. In this setup, GLAM-Swin-UNet yields 49.10% mIoU on ADE20K outperforming its vanilla Swin counterpart with at least 1.10% mIoU. GLAM-Swin-UperNet achieves 81.47 % mIoU on Cityscapes which is 1.58 % better than its Swin-Upernet counterpart.

Table 3.2: Comparison to state of the art methods on BCV.

Methods	Average Dice Score (%)
VNet [137]	68.81
U-Net [138]	76.85
Att-UNet [19]	77.77
R50-Deeplabv3+ [79]	75.73
TransUNet [107]	77.48
Swin-Unet [21]	79.13
TransClaw U-Net [139]	78.09
nnUNet (3D) [140]	86.99
nnFormer [2]	87.40
GLAM-nnFormer	**88.60**

Table 3.3: Comparison to state of the art methods on ADE20K and Cityscapes. All experiments are made or reported are with single-scale inference.

Method	Backone	ADE20K mIoU	Cityscapes mIoU
FCN [77]	ResNet-101	41.40	77.34
CCNet [141]	ResNet-101	43.71	79.45
DANet [142]	ResNet-101	43.64	80.47
UperNet [80]	ResNet-101	43.82	80.10
DNL [143]	ResNet-101	44.25	79.41
PSPNet [78]	ResNet-101	44.39	79.08
DeepLabV3+ [79]	ResNet-101	45.47	79.41
Trans2Seg [114]	PVT-S	42.60	-
FPN [114]	PVT-L	42.10	-
TNT [130]	TNT-S	43.60	-
SETR-PUP [144]	DeiT-L	46.34	79.21
Swin-UNet [21]	Swin-B	47.85	-
Swin-UperNet [17]	Swin-B	47.99	80.79
GLAM-Swin-UNet	Swin-B	**49.10**	-
GLAM-Swin-UperNet	Swin-B	48.44	**81.47**

3.4.3 Additional results

ADE20K In this additional experiment we use Multi Scales (MS) inference to evaluate the model and their extended GLAM version on ADE20K. As shown in 3.4, while MS inference improves the performances for all the methods, the GLAM models still outperform their baselines. Indeed, in this configuration, GLAM-Swin-UNet Base reach +1.55% on ADE20K and is still +0.93% higher than Swin-UNet Base.

Cityscapes We provide the same analysis on Cityscapes and compare the performances of Sinw-UNet Tiny and GLAM-Swin-UNet Tiny with and without MS inference as reported in 3.5. Again, GLAM-Swin-UNet Tiny outperforms Swin-UNet Tiny by 1% mIoU when trained over 40k epochs using MS inference. Moreover, we also give complementary results by providing performances with both models trained through 160k iterations. As can be seen in 3.5, the better performances of the GLAM model are stable as the GLAM-Swin-UNet outperforms its baseline by 0.80% mIoU and 1.09% mIoU with respectively SS and MS inference when trained through 160k epochs.

Synapse To explore more in depth the performance gain brought out by GLAM in Table 3 of the main paper, we show in 3.6 the segmentation results for the different organs of the dataset. The results are given for two baselines: TransUNet [107] and nnFormer [2] as well as for GLAM-nnFormer. We use the publicly available implementations provided by authors for both models[1,2]. The proposed

[1] https://github.com/Beckschen/TransUNet
[2] https://github.com/282857341/nnFormer

Table 3.4: **GLAM Improvements with Multi Scale inference on ADE20K.** Performances are evaluated with respect to mIoU for single scale inference (SS) and multiscales inference (MS).

Method	Size	SS	MS
Swin-UNet [21]	Tiny	42.75	44.72
GLAM-Swin-UNet	Tiny	**44.19**	**46.11**
Swin-UNet [21]	Base	47.85	49.72
GLAM-Swin-UNet	Base	**49.10**	**50.65**

Table 3.5: **GLAM Improvements with Multi Scale inference on Cityscapes.** Performances are evaluated with respect to mIoU for single scale inference (SS) and multiscales inference (MS).

Method	Size	SS	MS
Swin-UNet 40K [21]	Tiny	77.43	78.56
GLAM-Swin-UNet 40K	Tiny	**78.29**	**79.56**
Swin-UNet 160K [21]	Tiny	79.98	80.90
GLAM-Swin-UNet 160K	Tiny	**80.78**	**81.99**

GLAM-nnFormer sensibly outperform both baselines for all the classes except on the kidneys and the pancreas where the results are close to the standard nnFormer.

Table 3.6: Detailed per-organ comparison on the multi-organ Synapse dataset (Dice Score in %).

Methods	Aotra	Gallbladder	Kidnery(L)	Kidnery(R)	Liver	Pancreas	Spleen	Stomach
TransUNet [107]	87.23	63.16	81.87	77.02	94.08	55.86	85.08	75.62
nnFormer [2]	89.81	63.18	93.78	**94.58**	96.19	**83.16**	95.76	86.14
GLAM-nnFormer	**90.10**	**65.81**	**93.92**	**94.56**	**96.74**	82.91	**96.49**	**88.20**

3.4.4 Model Analysis

In this part, we analyze various important aspects of GLAM.

Number of Global Tokens. The number of global tokens directly influences the capacity of GLAM to model global interactions between the windows. Fig. 3.4 shows the impact of this hyper-parameter on segmentation performances. We can see that using more global tokens improves performance. However, it also increases the number of parameters and memory cost which forces a trade-off. We keep a reasonable value of 10 global tokens, which gives an important performance boost of +1.4pts in both the tiny and base versions of the Swin-UNet model.

Impact of NLU. GLAM improves context modeling in multi-resolution transformers thanks to global attention and Non-Local Upsampling (NLU). Table 3.7 provides an ablation study of these two components. We can see that NLU gives an improvement of 0.45pt compared to the original Swin-

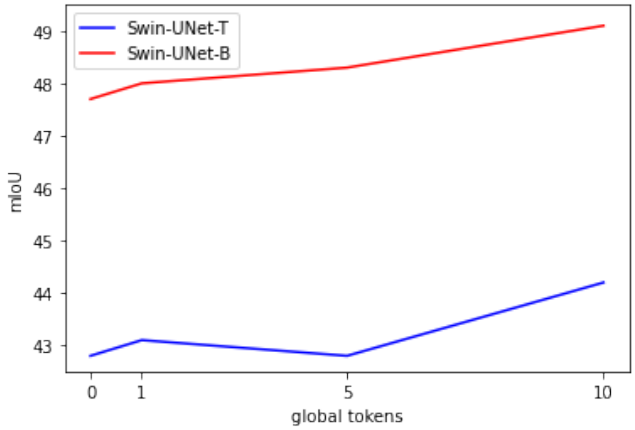


Figure 3.4: Impact of the number of global tokens on performance (mIoU) using ADE20k.

Table 3.7: Impact of the NLU and the GLAM transformer on a tiny Swin-UNet, 10 global tokens, on ADE20k.

Method	NLU	GLAM	mIoU
Swin-UNet-T			42.75
Swin-UNet-T	✓		43.20
Swin-UNet-T	✓	✓	**44.20**

UNet that uses a patch expansion operation for upsampling. GLAM brings another large improvement for a total gain of +1.44pts compared to the baseline.

Long-range interaction. To highlight the impact of G-MSA, Table 3.8 shows the performances of GLAM backbones using only a W-MSA step but no G-MSA. GLAM backbones show consistent gains compared to their counterparts without G-MSA. This ablation highlights the crucial role of this step to leverage long-range interactions and that the performance gains made by GLAM can not only be explained by the parameter overhead.

Table 3.8: Impact of G-MSA phase on GLAM transformer on different model, 10 global tokens, on ADE20k. GLAM-nogmsa is GLAM without the G-MSA phase.

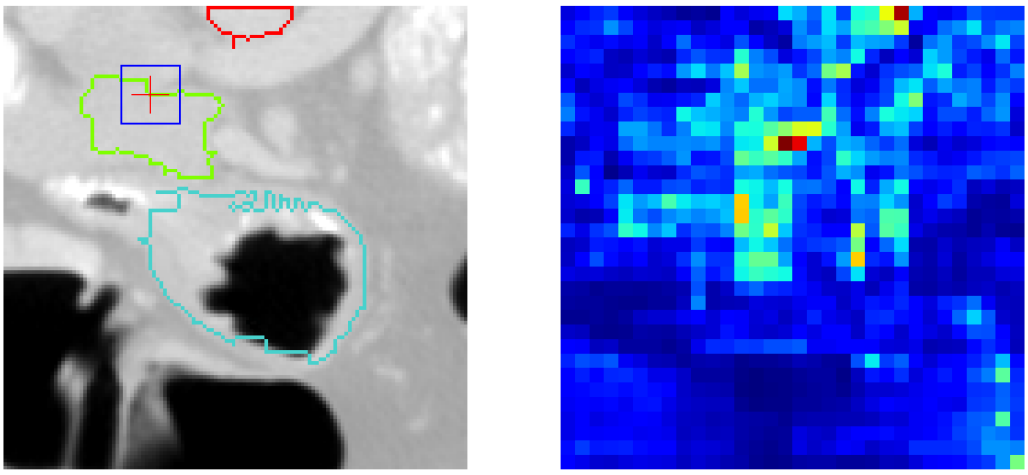
Method	mIoU
GLAM-nogmsa-Swin-UNet B	47.90
GLAM-Swin-UNet B	**49.10**
GLAM-nogmsa-Swin-UperNet B	47.95
GLAM-Swin-UperNet B	**48.44**

Parameter and FLOPs overhead. The overhead due to the global tokens is controlled and proportional to the number of GLAM transformer blocks. This overhead brings higher performance gains

than increasing the backbone size which validates the model architecture. Table 3.9 illustrates that the GLAM-Swin Base backbones show superior efficiency compared to their vanilla Large counterpart with a superior mIoU increase with respect to additional learnable parameters. The same analysis can be done with FLOPs overhead with a higher mIoU increase per extra-FLOP for GLAM-Swin Base compared to Swin Large.

Table 3.9: Analysis of the relative mIoU increase with respect to extra learnable parameters and FLOPs compared to the standard Base and Large backbones.

backbone	#param.	↑ rel. mIoU / #param $\times 10^{-2}$	FLOPs	↑ rel. mIoU / FLOPs $\times 10^{-2}$
Swin-UperNet B	121	0	81G	0
Swin-UperNet L	234	0.4	180G	0.4
GLAM-Swin-UperNet B	197	**0.6**	99G	**2.5**



Input image GLAM global attention

Figure 3.5: **Averaged GLAM attention map in 3D.** The information inside the blue window is ambiguous. To segment the voxel at the red cross, the model leverages long-range dependencies including neighbor organs. The pancreas is in green, the aorta in red, and the stomach in blue.

Global token merging strategy. Here, we study the importance of how the global tokens between different windows are merged: with the GLAM transformer, we use a global self-attention (G-MSA) mechanism. We compare G-MSA with an averaging and a random permutation strategy. We can see in Table 3.10 that G-MSA is largely superior to the two other options. This validates the usefulness of the G-MSA step, which enable indirect modeling of full-range interactions between visual region when applied after W-MSA.

Table 3.10: Global token merging (tiny Swin-Unet, ADE20k).

Merging strategy	mIoU
Random permutation	43.2
Average	43.7
G-MSA Merging	**44.2**

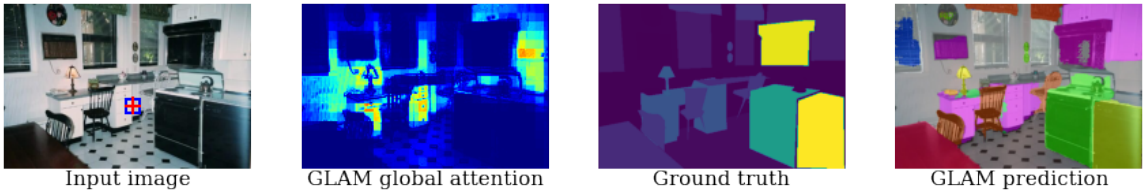
3.4.5 Visualizations.

ADE20K and Cityscapes. Fig. 3.6 shows qualitative visualizations of the GLAM method. In Fig. 3.6a), we show GLAM attention maps for the highest resolution feature maps of a GLAM Swin-Unet model. Echoing observations in Fig. 3.1 in Cityscape, we can see that GLAM can model full-range interactions in this spatially fine layer. This enables to exploit spatial relationships with other important structures (*e.g.* other sofas, arcades), which is not possible with the baseline Swin-Unet due to its limited window attention. We can notice the relevance of the GLAM segmentation. Furthermore, Fig. 3.5 shows the GLAM attention averaged over the axial direction for the red cross (pancreas). We can see that long-range dependencies are involved, with a much larger spatial extent than the local window (in blue), where attention is given to neighboring organs (stomach and aorta). The full context is crucial to properly segment complex organs with visual local ambiguities such as the pancreas.

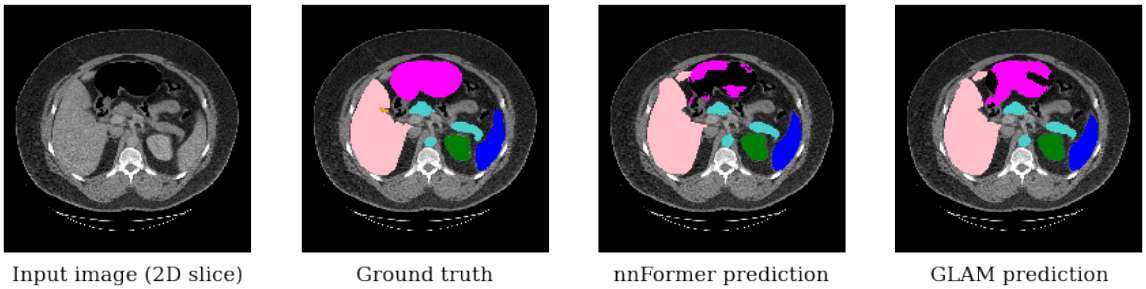
In 3.8 and 3.9, we select some representative images of the GLAM-Swin-Unet and GLAM-Swin-Upernet. We provide attention maps for the lowest hierarchy as well as the generated segmentation map for the GLAM models. The attention is computed with respect to a global token associated to the 7×7 blue window plotted in the image (not to the scale). For the first stage of the model, the patch size is 4×4 patches and thus the dimension of the window is 28×28 pixels. We see that the model manages to detect long-range interactions directly in high-resolution feature maps without being limited by the small window size. Attention is paid mostly between elements of the same class: vegetation in 3.8, chairs or sky in 3.9 but also to salient elements such as corners or edges and semantic ones such as cars and pedestrians.

We provide another comparison on ADE20K in Fig. 3.7 below. Again, we can notice that Swin's attention is limited to the small blue region. In contrast, GLAM can compute a global attention map at high-resolution thanks to the G-MSA module, providing both accurate spatial information and global context.

BCV. In Fig. 3.6b), we show segmentation results of GLAM-nn-Former for 3D medical image segmentation. We show the results on a given 2D slice. We can notice that GLAM nn-Former is



a) Segmentation results and global attention of GLAM on ADE20K.



b) Segmentation results on BCV.

Figure 3.6: Qualitative visualisations of GLAM. We show the ability of GLAM to model full contextual information in high-resolution feature maps on ADE20K (first row), and the ability of GLAM-nnFormer to accurately segment the stomach (in pink).

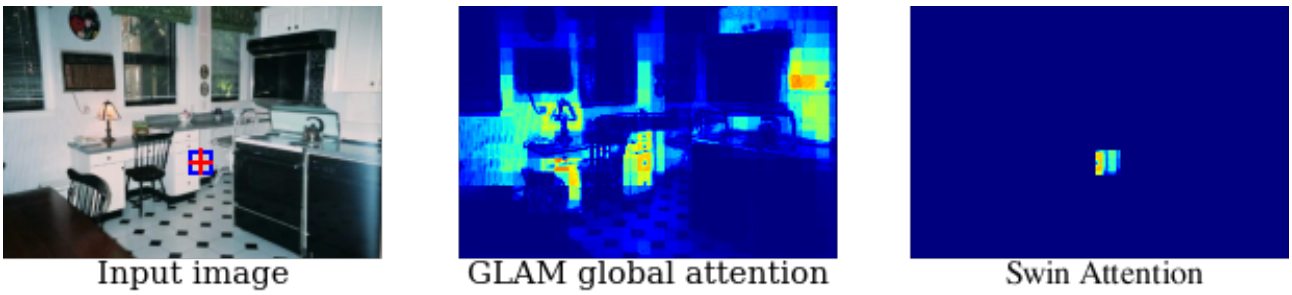


Figure 3.7: Global attention of GLAM compared to vanilla Swin on ADE20K.

qualitatively much better at segmenting the stomach (in pink) than nn-Former. This can be explained by the global interactions of our model, which enables it to better represent specific interactions between organs.

In 3.10, we present more segmentation results on 3D medical images and provide a qualitative analysis of the performances between nnFormer and GLAM-nnFormer. The GLAM model manages to retrieve better segmentation of the liver (pink) and the stomach (purple). The memory effect of the global tokens manages to limit the error due to the inference on 3D crops which is well illustrated on the liver reconstruction.

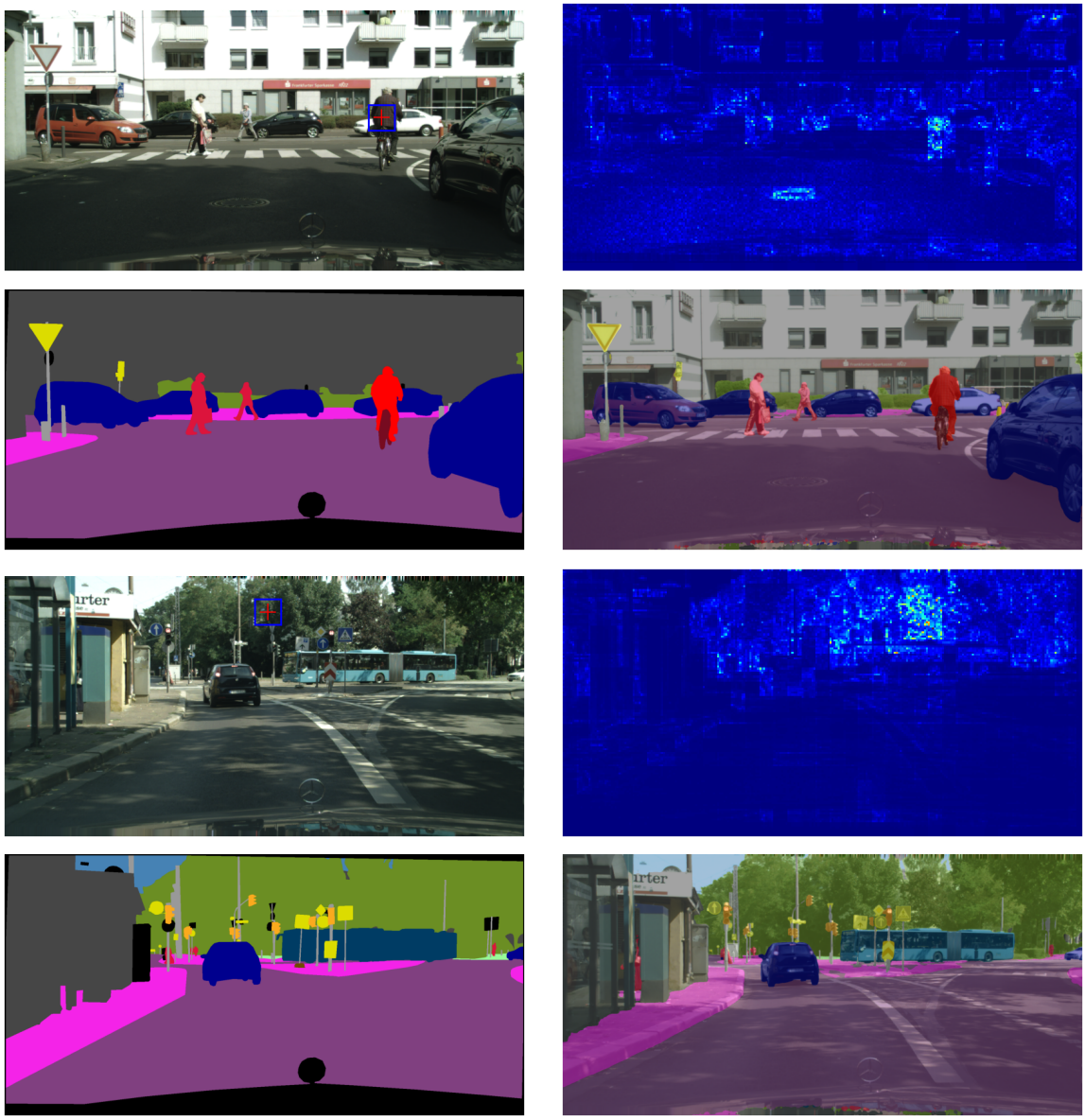


Figure 3.8: **Qualitative results of GLAM incorporated to Swin-upernet on Cityscapes** For two test images, we show from top-left to bottom-right : the image, the global attention map with respect to the blue window, the ground truth and the predicted segmentation.

3.5 Conclusion

This chapter introduces GLAM, a method for modeling full contextual interactions in multi-resolution transformer-based models. The GLAM transformer leverages learnable global tokens at each resolution level of the model, which allows a complete interaction of the tokens across the image

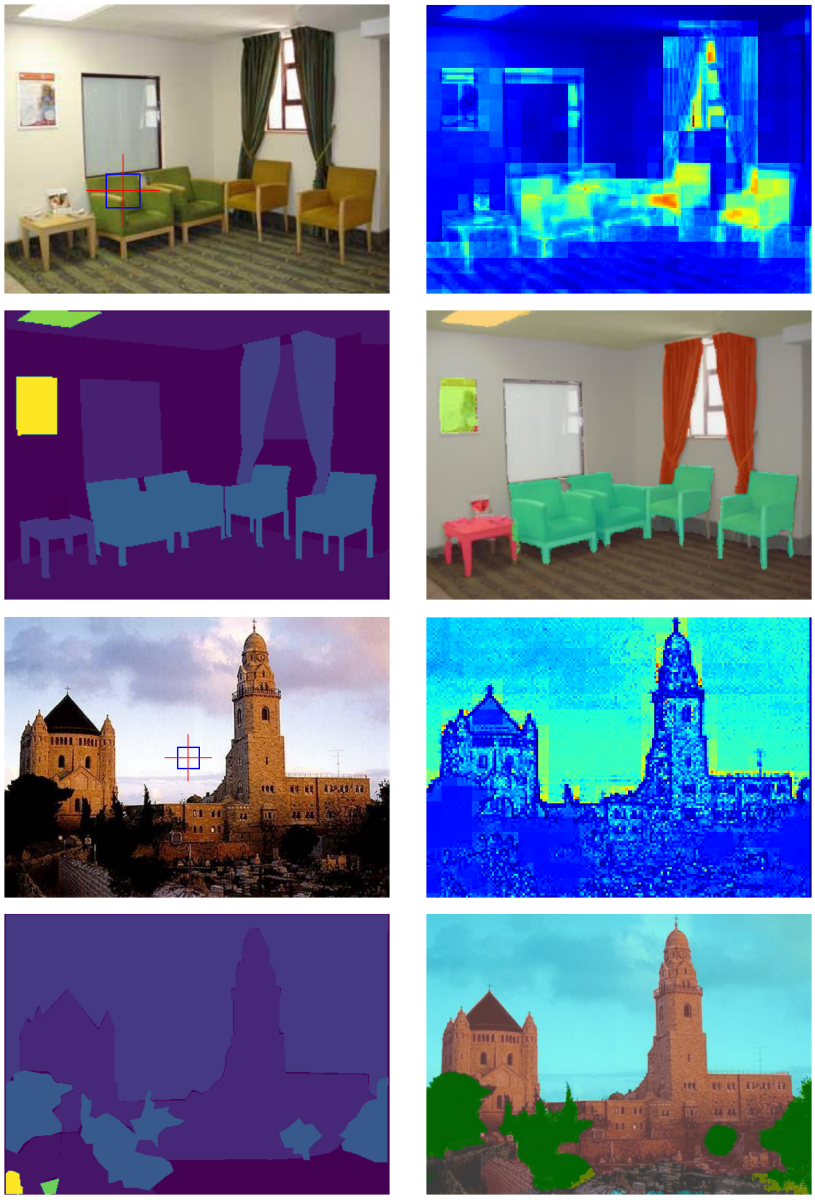


Figure 3.9: **Qualitative results of GLAM incorporated to Swin-UNet on ADE20K** For two test images, we show from top-left to bottom-right : the image, the global attention map with respect to the blue window, the ground truth and the predicted segmentation.

regions. Experiments show the large and consistent gain of GLAM when incorporated into several multi-resolution transformers (Swin-UNet, nn-Former, Swin) on diverse natural, panoptic or medical datasets.

Nevertheless, the typical size of 3D medical images doesn't allow to directly segment the full

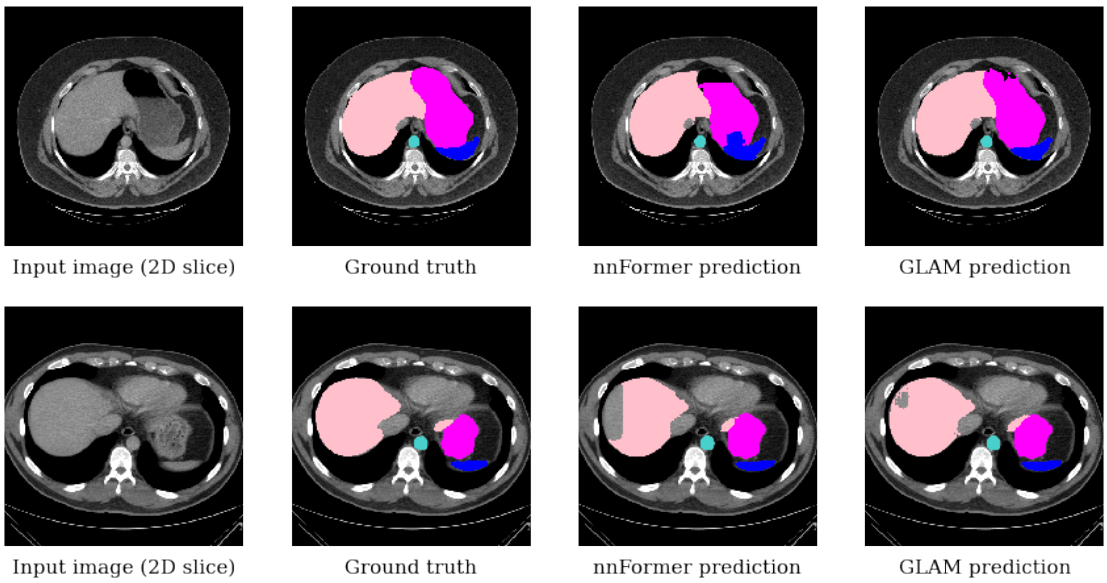


Figure 3.10: **Qualitative results of GLAM Incorporated to nnFormer on BCV.** The observed organs are the liver (pink), the stomach (purple), the aorta (cyan) and the spleen (blue).

volume. A common strategy is to train the network on randomly cropped patches. If this strategy ensures relatively good performances and generalization, it strongly restricts the context of the model and thus the scale of the information to grasp. For 3D medical image segmentation, GLAM is able to model full interaction in the cropped patch input but is still unable to model interactions beyond this cropped patch. These out-of-range interactions in the full original volume are not exploited. To deal with this issue, in the next chapter, we present an extension of GLAM that is able to indirectly model out-of-range interactions.

Chapter 4

LORI: Long and Out of Range Interaction transformer module for 3D medical image segmentation

Contents

4.1	Introduction		**97**
4.2	Indirect attention modeling		**98**
4.3	The Full resolutIoN mEmory transformer (FINE)		**99**
4.4	Long-and-Out-of-Range Interaction method (LORI)		**101**
4.5	Experiments		**106**
	4.5.1	Datasets	106
	4.5.2	Implementation details	106
	4.5.3	Backbones	107
4.6	Results		**108**
	4.6.1	Preliminary results	108
	4.6.2	Results	109
	4.6.3	LORI Model analysis	111
4.7	Conclusion		**115**

Chapter summary

We saw in the last chapter that indirect attention is a promising tool to model long-range interaction even at high-resolution. This concept extends well to very large medical volumes which are typically

trained on random cropping. If this training strategy allows scalable training on high-resolution medical volumes, the attention of the model is restricted to the cropped patch boundaries. This means that with this strategy the model lost what we call out-of-range interactions. In this chapter, we introduced a novel transformer-based method to effectively model high-resolution interactions inside the cropped patch and still capture out-of-range interactions from the large anatomical structure by building upon the concept of indirect attention introduced in GLAM in Chapter 3. We introduce a new transformer method that aims to address the aforementioned challenges by facilitating the incorporation of long-and-out-of-range dependencies in medical segmentation models. This method incorporates global tokens that serve to propagate global representations between the different regions of the image. We provide two variants of this method: FINE (Full resolutIoN mEmory transformer), a full transformer architecture that works as a preliminary proof of concept, and LORI (Long and Out-of-Range Interaction transformer) which is a generic module allowing it to be seamlessly integrated into existing models such as nnUNet. FINE is a proof of concept extending GLAM method by using two levels of global tokens for long-range and out-of-range interactions modeling while LORI utilizes only one type of global token to model all interactions simultaneously, resulting in enhanced efficiency. We performed preliminary experiments on BCV with FINE showing its relevance, and extensive experimental evaluations with LORI on three distinct datasets: two 3D CT multi-organs segmentation datasets and one 3D ultrasound image dataset for liver and vessel segmentation. The results obtained from these evaluations demonstrate the consistent enhancement in segmentation performance achieved by LORI. Notably, LORI exhibited superior performance across multiple multi-class high-resolution 3D image datasets, irrespective of the different modalities involved.

4.1 Introduction

Nowadays, most existing DL based segmentation methods [34, 14, 35, 36] can not handle full 3D medical images and are limited to processing sub-regions of the input image *ie.* cropped patches. Contrary to 2D slices, those patches enable to preserve the 3D nature of the input, while keeping the original resolution of the volume and maintaining all the fine-grained details. However, this approach is not free of drawbacks. Indeed, patches are processed independently, leading to a dramatic loss of context: information outside the crop *ie.* out-of-range information, cannot be used in the prediction and is lost. Consequently, when facing challenging segmentation scenarios, *e.g.*, intricate organs or noisy data, the models often struggle to produce accurate segmentation. As illustrated in Fig. 4.1, the input patch, represented by the green square, encompasses only a limited portion of the original image. Consequently, the amount of available information within this patch is insufficient for an accurate segmentation of the kidney, depicted in yellow. The main objective of this chapter is to address this issue by using a new method to model out-of-range interactions.

In this chapter, we generalize the concept of global tokens as a pivot to spread multi-scale information in the attention. Compared to standard self-attention applied on raw 3D volume, this approach provides a way to model global context while maintaining memory usage and computational cost under control. Thus, we present the Full resolutIoN mEmory transformer (FINE) and the Long and Out-of-Range Interaction transformer (LORI).

In Fig. 4.1, LORI shows its capacity to effectively leverage both long-range and out-of-range information modeling meaning that both high-resolution information inside the crop and in the full volume are processed. As represented by the attention map, each pixel can indirectly attend to any other pixel of the original volume, even outside the input patch. Moreover, the global tokens can learn by them-self strong positional information of the region they are assigned to. In this way, it enables the model to accurately segment the pixel indicated by the red cross, by benefiting from a larger context.

The main contributions of this chapter are as follows:

- We introduced FINE, which serves as a preliminary work for LORI. FINE is a proof of concept for modeling of long and out-of-range interactions through global tokens in the context of 3D medical image segmentation. FINE uses one type of global tokens for long-range information and another one for out-of-range information. Then FINE uses two Transformer modules to create the information indirection.

- We propose LORI, an efficient module for modeling long and out-of-range interactions through global tokens, enabling the effective dissemination of information from the entire original volume to all feature maps, bringing consistent gains to the performance of state-of-the-art

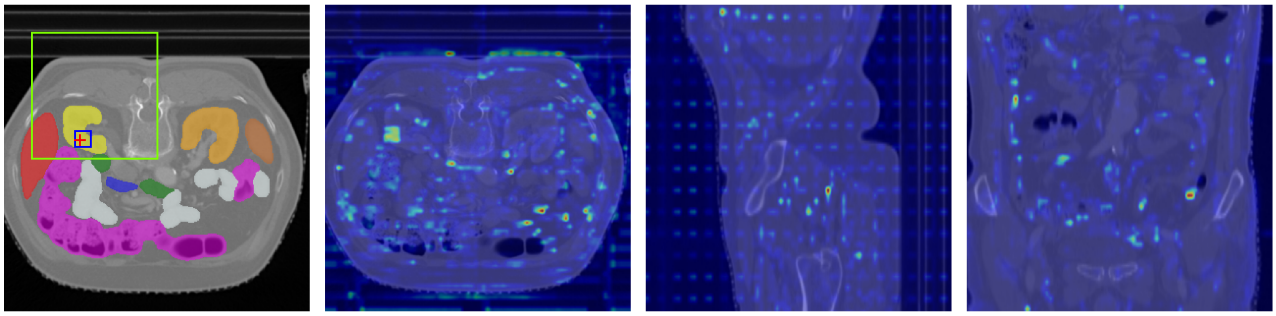


Figure 4.1: Visualization of LORI's Indirect Attention on the WORD Dataset. The red cross indicates the focal pixel, with the attention map highlighting its corresponding attention. Unlike the limitation imposed by the window size (blue square), LORI efficiently computes long-range attention across the entire input cropped image (green square). Furthermore, LORI effectively captures out-of-range attention from the complete volume, as demonstrated through visualizations on axial, sagittal, and coronal planes. These visualizations demonstrate LORI's capacity to harness information from diverse spatial dimensions, underscoring its potential in significantly enhancing 3D medical image segmentation.

methods. LORI uses only one type of global tokens to model all interactions and thus uses only one Transformer making it more efficient as it requires less parameters.

- We demonstrate the versatility of LORI through its integration into various state-of-the-art segmentation architectures: CNNs with nnUNet[34], Transformers with 3D Swin-UNet[35], and hybrid such as CoTr [14].

- We show the superiority of LORI compared to state-of-the-art models on two significant 3D CT datasets for multi-organ segmentation, as well as a private dataset for liver and vessel segmentation in 3D ultrasound images.

4.2 Indirect attention modeling

We aim to model large-scale and high-resolution interactions by generalizing the concept of information indirection through global tokens. This necessitates the identification of three levels of information. Firstly, at the window level, we want to preserve fine-grained details and thus compute full quadratic self-attention. The second level corresponds to the cropped patch extension over which we aim to propagate local windows information. Lastly, at the global level, information is derived from the overall volume which describes the high-level structures within the image. Ideally, we would

like the information to cascade from the global level to the window level. To circulate information, we need to subdivide each level into sub-regions to be paired with dedicated global tokens.

Generic Notations. We will present here the notations needed for the methods:

- $H \times W \times D$ the full volume dimensions
- $H_r \times W_r \times D_r$ the dimension of the large volume sub-regions
- n_g the number of global tokens associated with each of the sub-regions
- $N_c = \frac{H \times W \times D}{H_r \times W_r \times D_r}$ the number of sub-regions in a large volume
- N_w the number of windows in the patch cropped for training
- $\boldsymbol{x} \in \mathbb{R}^{H_r \times W_r \times D_r}$ the patch cropped from the large volume to be processed by the neural network
- N_p the number of visual tokens paired with each window $\boldsymbol{v}_k$
- $\{\boldsymbol{v}_k\}_{1 \leq k \leq N_w} \in \mathbb{R}^{N_p \times d}$ the set of visual token sequences related to the $k-th$ window of the input $\boldsymbol{x}$
- Visual tokens *ie.* inside a window will be in blue and global tokens in red
- **NB:** To ease the computation and restrict the number of global tokens, we chose the sub-regions and the windows to be non-overlapping. Moreover, the cropped patch and the volume sub-regions dimensions share the same dimensions.

4.3 The Full resolutIoN mEmory transformer (FINE)

We introduce the specific following notations for this sub-section only:

- $\boldsymbol{o} \in \mathbb{R}^{(N_w \cdot N_o) \times d}$: sequence of N_w window-level global tokens
- $\boldsymbol{W} \in \mathbb{R}^{(N_c \cdot N) \times d}$: sequence of N_c volume-level global tokens

The core idea of FINE is to introduce global tokens to enable full-range interactions between all voxels at all resolution levels with random cropping. We introduce global tokens at two levels. First, we add specific global tokens to the sequence of visual tokens from each window. We chose to call them window tokens to avoid any confusion. The second level of global tokens will be used to keep track of the observed part of the volume. We will refer to these dedicated global tokens as volume

tokens. These volume tokens are associated with each element of the grid of sub-regions covering the entire volume and are called by the transformers when performing the segmentation of a cropped patch. As can be seen in Fig.4.2, the volume tokens induce a positional encoding learned over the entire volume.

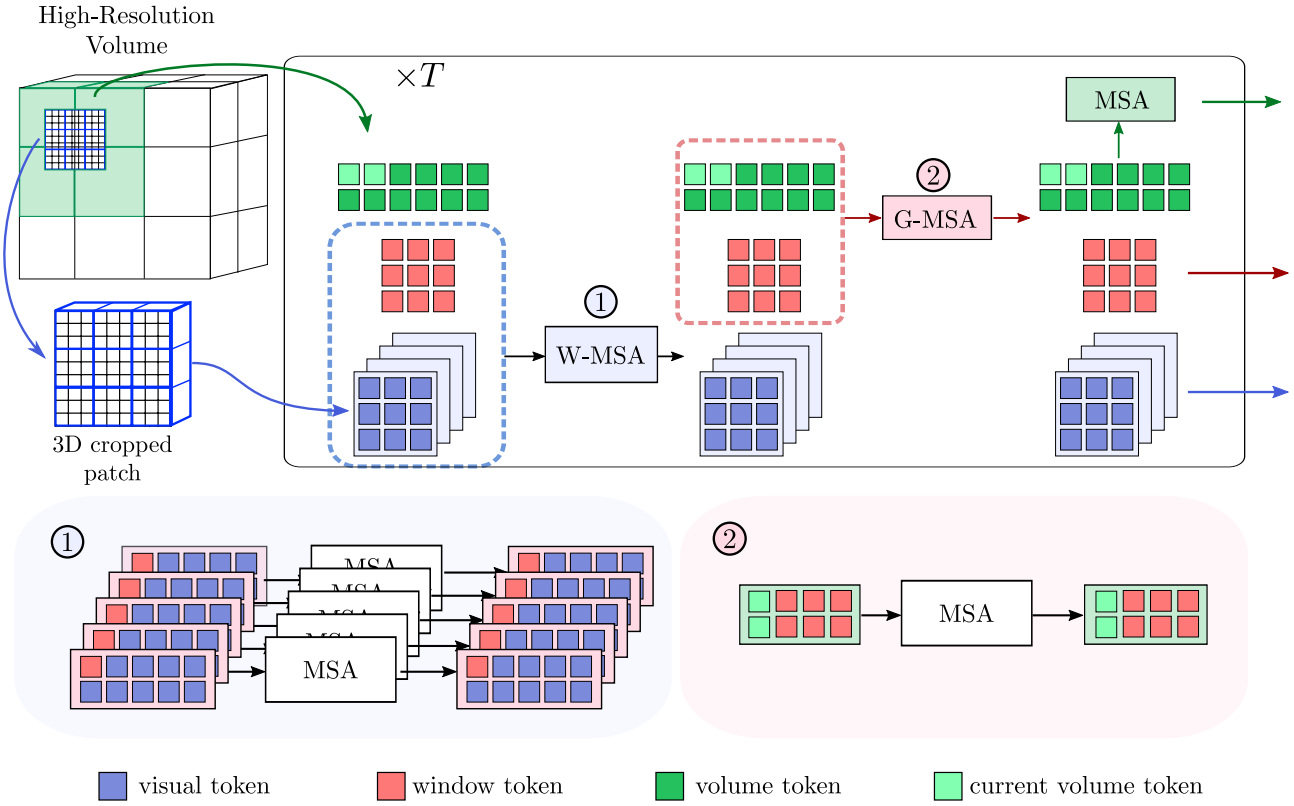


Figure 4.2: To segment the cropped patch in blue and model global context, two levels of memory tokens are introduced: window (red) and volume (green) tokens. First, the blue crop is divided into windows over which Multi-head Self-Attention (MSA) is performed in parallel. For each window, the sequence of visual tokens (blue) is augmented with a specific window token. Second, the local information embedded into each window token is shared between all window tokens and volume tokens intersecting with the crop (light green). Finally, high-level information is shared between all volume tokens).

In Fig.4.2, volume, window, and visual tokens are indicated in green, red, and blue respectively. First, MSA is performed for each window over the merged sequence of visual and window tokens. Given a sequence of visual tokens *ie.* small patches, W-MSA consists of computing the MSA in parallel for all windows composing the sequence. G-MSA is performed over the merged sequence of all window tokens and corresponding volume tokens to grasp long-range dependencies in the input patch. G-MSA involves only the volume token corresponding to sub-volumes intersecting with the input patch $\boldsymbol{x}$. Finally, full-resolution attention is achieved by applying MSA over the sequence

of volume tokens. Formally, the l-th FINE-transformer bloc is composed of the following three operations:

$$\begin{aligned} [\boldsymbol{v}^l, \boldsymbol{o}^l] &= \text{W-MSA}([\boldsymbol{v}^{l-1}, \boldsymbol{o}^{l-1}]), \\ [\boldsymbol{o}^l, \hat{\boldsymbol{w}}^l] &= \text{G-MSA}([\boldsymbol{o}^{l-1}, \boldsymbol{w}^{l-1}]), \\ \boldsymbol{W}^l &= \text{MSA}(\hat{\boldsymbol{W}}^l). \end{aligned} \tag{4.1}$$

$\boldsymbol{w}$ denotes the volume tokens corresponding to sub-volumes with a non-null intersection with $\boldsymbol{x}$ and $[\boldsymbol{a}, \boldsymbol{b}]$ stands for the concatenation of $\boldsymbol{a}$ and $\boldsymbol{b}$ along the first dimension.

4.4 Long-and-Out-of-Range Interaction method (LORI)

We introduce the specific following notations for this sub-section only:

- $\boldsymbol{G} \in \mathbb{R}^{N_g \times d}$: sequence of all grounded global tokens
- N_r the number of sub-regions overlapping with $\boldsymbol{x}$
- $\boldsymbol{g} \in \mathbb{R}^{N_r \times d}$: sequence of grounded global tokens associated with the regions overlapping with $\boldsymbol{x}$
- $\{\boldsymbol{z}_k^0\}$ with $\boldsymbol{z}_k^0 = [\boldsymbol{v}_k^0, \boldsymbol{g}^0]$ the merged sequences of visual tokens for the window k and the global tokens related to the cropped area.

As illustrated in Fig. 4.3, LORI uses global tokens grounded to regions of the input volume. During a forward pass, global tokens related to regions overlapping with the cropped patch are injected into the model (Fig. 4.3 (a)). Through LORI, these global tokens circulate information in the patch (Fig. 4.3 (b)) and infuse out-of-range information (Fig. 4.3 (c)).

Grounded Long-Range Interaction Global tokens behave as local representations of specific anatomical parts. The fact that all of these parts are not available when segmenting a given crop is a challenging configuration to learn these representations. Rather, than learning two separate sets of global tokens dedicated to extract useful representations of the underlying structure and propagate high-resolution information between windows, we use only one level of global tokens grounded to each sub-region of the volume. These global tokens are learned asynchronously by updating only the ones associated with the regions overlapping with the training patches $\boldsymbol{x}$.

Out-of-range interaction Injecting grounded global tokens into the GLAM module gives the model the ability to use learned representations of the regions surrounding the patch. This information

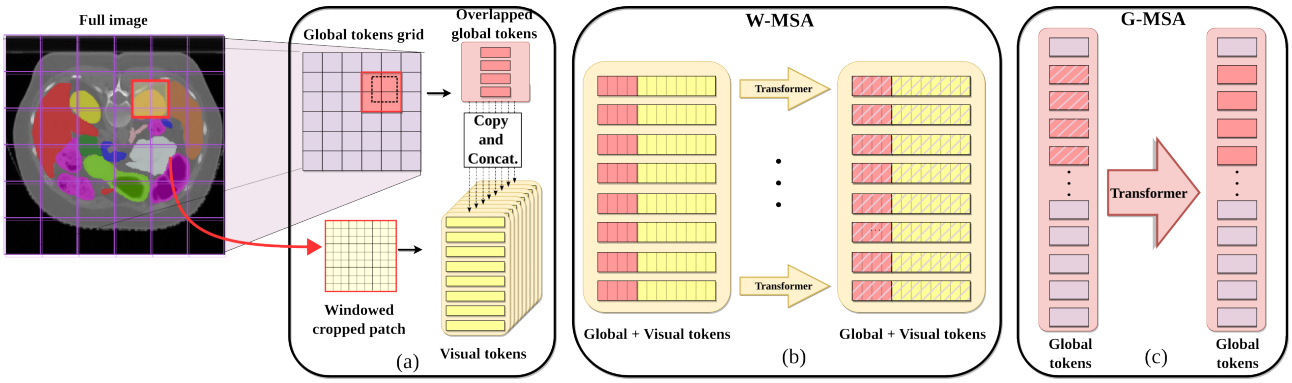


Figure 4.3: Overview of the LORI module: (a) shows the whole set of global tokens of LORI (purple tokens), each one associated with a region of the image. When a cropped patch is selected into the full-size image (red square), the associated global tokens are selected (red tokens). The set of visual tokens (yellow tokens) is also selected and reorganized into a subset of windows. (b) shows the first attention step of LORI where the selected global tokens are duplicated to form the global feature map and to be associated with each window. A window transformer (W-MSA) is applied to this feature map to share information between the global feature map and the visual tokens. (c) shows the second attention step of LORI done by a classic transformer (G-MSA) on the features composed of the global feature map which captured information from each window and the set of all global tokens. This step lets the global feature map share information between all windows and incorporates information from the full-size image.

allows the model to align the visual tokens of the input with the learned high-level representations thus improving their relevance during training. To go one step further, by chaining multiple W-MSA and G-MSA operations, LORI uses not only surrounding information but the representations of the total underlying structure allowing the model to indirectly grasp interactions beyond the observed patch. LORI's complete sequence of operations is given by a slight modification of the GLAM introduced previously and the l-th block of the module has the following form:

$$\begin{aligned} \hat{\boldsymbol{z}}^{l} &= \text{W-MSA}(\boldsymbol{z}^{l-1}), \\ \boldsymbol{g}^{l} &= \text{G-MSA}\left(\left[\boldsymbol{G}^{T}, \hat{\boldsymbol{g}}^{l^{T}}\right]^{T}\right), \\ \forall k \in \{1, \cdots, N_w\}\ \boldsymbol{z}_k^{l} &= \left[\boldsymbol{v}_k^{l^{T}}, \boldsymbol{g}_k^{l^{T}}\right]^{T}. \end{aligned} \tag{4.2}$$

LORI possesses the capability to replicate the information exchange among windows performed by long-range interaction modeling. Additionally, through G-MSA, the attention between the selected global tokens $\boldsymbol{g}^{l}$ and all the global tokens $\boldsymbol{G}^{l}$ is evaluated, and information from the entire volume is indirectly shared. Subsequently, in block $l+1$, the out-of-range information gathered by $\boldsymbol{g}^{l}$ is

transmitted to the visual tokens of the cropped region $\mathbf{v}^{l+1}$, thereby sharing the captured information.

Upon comparing LORI to FINE [40], it becomes evident that LORI exhibits a more streamlined architecture, employing only two multi-head self-attention operations: one for window attention and another for long and out-of-range attention, as expressed by Eq. 4.2. In contrast, FINE utilizes three separate steps, partitioning the **G-MSA** into two parts. This dissimilarity grants LORI enhanced efficiency by requiring fewer parameters and operations for establishing the information indirection. Furthermore, LORI enables easier propagation of out-of-range information, necessitating only two modules in the layer to transfer the information from global tokens to visual tokens, in contrast to FINE's three steps, resulting in a lighter overall architecture.

Indirect LORI attention As depicted in Fig. **??**(c), LORI facilitates the indirect computation of attention across the entire input volume. The attention between the global tokens (depicted on the left) that capture long-range and out-of-range information is integrated with the local visual tokens, allowing for the indirect propagation of information across all visual tokens and thus indirect full attention. To observe this attention and thus better understand the method, we can develop the indirect attention formula based on the LORI's operations described earlier. The following section will show how to compute the indirect attention of LORI mechanism.

Let note: i, j and l the i^{th} cropped patch, j^{th} window and l^{th} LORI's block.

The attention matrix of LORI's W-MSA for crop i, window j, and block l can be denoted as $\mathbf{A}^l_{ij}$. Similarly, the attention matrix of LORI's global tokens MSA for crop i and block l can be represented as $\mathbf{B}^l_i$. We describe these matrices as:

$$\mathbf{A}^l_{ij} = \begin{bmatrix} \mathbf{A}^l_{ij,vv} & \mathbf{A}^l_{ij,vg} \\ \mathbf{A}^l_{ij,gv} & \mathbf{A}^l_{ij,gg} \end{bmatrix} \quad \mathbf{B}^l_i = \begin{bmatrix} \mathbf{B}^l_{i,GG} & \mathbf{B}^l_{i,Gg} \\ \mathbf{B}^l_{i,gG} & \mathbf{B}^l_{i,gg} \end{bmatrix} \tag{4.3}$$

We define $\mathbf{A}_{ij,xy}$ or $\mathbf{B}_{i,xy}$ the sub-matrices of the attention between a set of token x as queries and an other set y as keys and values.

In the following proof, we aim to obtain the indirect global attention matrix for the visual tokens of window j, crop i, and block l. To accomplish this, we consider the indirect cumulative contribution of all tokens in the construction of $\mathbf{v}^{l+1}_{ij}$ which is indirectly based on all the visual tokens $\mathbf{v}^{l-1}_{cw}$ for all cropped patch c and window w.

Thanks to Eq. 3.2 and Eq. 4.3 we can describe the composition of each token with their associated attention matrix like this:

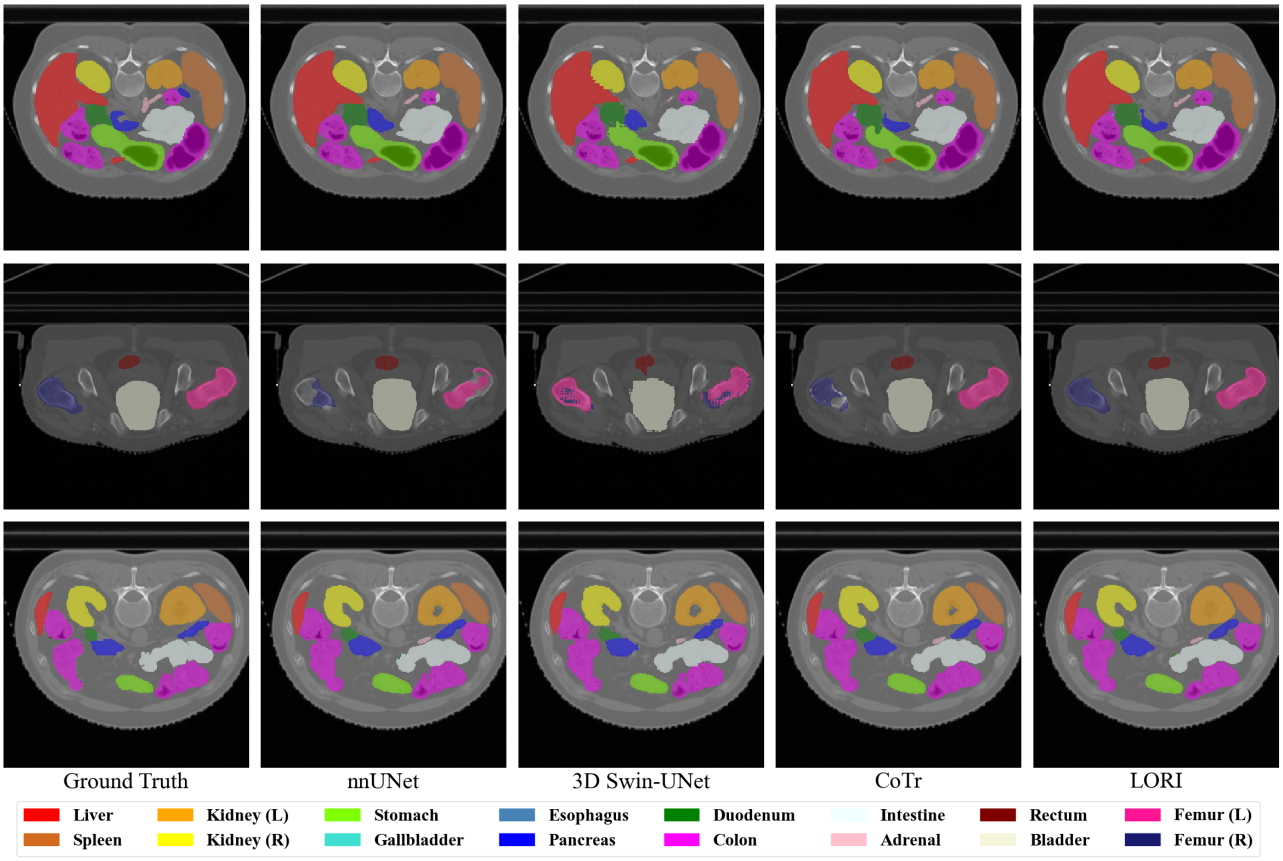


Figure 4.4: Visualisation of segmentation masks produced by the different methods on the test set of WORD dataset.

$$\begin{cases} \mathbf{v}_{ij}^{l} = \mathbf{A}_{ij,vv}^{l} \quad \mathbf{v}_{ij}^{l-1} + \mathbf{A}_{ij,vg}^{l} \quad \mathbf{g}_{ij}^{l-1} \\ \hat{\mathbf{g}}_{ij}^{l} = \mathbf{A}_{ij,gv}^{l} \quad \mathbf{v}_{ij}^{l-1} + \mathbf{A}_{ij,gg}^{l} \quad \mathbf{g}_{ij}^{l-1} \\ \mathbf{G}_{i}^{l} = \mathbf{B}_{i,GG}^{l} \quad \mathbf{G}_{i}^{l-1} + \mathbf{B}_{i,Gg}^{l} \quad \hat{\mathbf{g}}_{i}^{l} \\ \mathbf{g}_{ij}^{l} = \mathbf{B}_{i,gG,j}^{l} \quad \mathbf{G}_{i}^{l-1} + \mathbf{B}_{i,gg,j}^{l} \quad \hat{\mathbf{g}}_{i}^{l} \end{cases} \tag{4.4}$$

Then, using Eq. 4.4 we can develop $\mathbf{v}_{ij}^{l+1}$:

$$\begin{aligned} \mathbf{v}_{ij}^{l+1} &= \mathbf{A}_{ij,vv}^{l+1}\mathbf{v}_{ij}^{l} + \mathbf{A}_{ij,vg}^{l+1}\mathbf{g}_{ij}^{l} \\ &= \mathbf{A}_{ij,vv}^{l+1}(\mathbf{A}_{ij,vv}^{l}\mathbf{v}_{ij}^{l-1} + \mathbf{A}_{ij,vg}^{l}\mathbf{g}_{ij}^{l-1}) + \mathbf{A}_{ij,vg}^{l+1}(\mathbf{B}_{i,gg,j}^{l}\hat{\mathbf{g}}_{i}^{l} + \mathbf{B}_{i,gG,j}^{l}\mathbf{G}_{i}^{l-1}) \\ &= \mathbf{A}_{ij,vv}^{l+1}\mathbf{A}_{ij,vv}^{l}\mathbf{v}_{ij}^{l-1} + \mathbf{A}_{ij,vg}^{l+1}\sum_{w}^{N_w}(\mathbf{B}_{i,gg,jw}^{l}\hat{\mathbf{g}}_{iw}^{l}) + \mathbf{A}_{ij,vg}^{l+1}\sum_{c}^{N_c}(\mathbf{B}_{i,gG,jc}^{l}\mathbf{G}_{i,c}^{l-1}) + Cst \\ &= \mathbf{A}_{ij,vv}^{l+1}\mathbf{A}_{ij,vv}^{l}\mathbf{v}_{ij}^{l-1} + \mathbf{A}_{ij,vg}^{l+1}\sum_{w}^{N_w}(\mathbf{B}_{i,gg,jw}^{l}\mathbf{A}_{iw,gv}^{l}\mathbf{v}_{iw}^{l-1}) + \mathbf{A}_{ij,vg}^{l+1}\sum_{c}^{N_c}(\mathbf{B}_{i,gG,jc}^{l}\mathbf{G}_{i,c}^{l-1}) + Cst \end{aligned} \tag{4.5}$$

Now, to visualize the indirect attention that contributes to construct $\mathbf{v}_{ij}^{l+1}$ outside of the cropped patch i, we pose that $\mathbf{G}_{i,c}^{l-1} = \mathbf{G}_{c,c}^{l} \forall c \in [1,...,N_c]$. We can do this assumption because $\mathbf{G}_{i,c}^{l-1}$ represent an estimation of the information contained in the cropped patch c and $\mathbf{G}_{c,c}^{l}$ is the global tokens that represent the most the cropped patch c as it has captured information from its visual tokens. So we can write:

$$\begin{aligned}
\mathbf{v}_{ij}^{l+1} =& \mathbf{A}_{ij,vv}^{l+1}\mathbf{A}_{ij,vv}^{l}\mathbf{v}_{ij}^{l-1} + \mathbf{A}_{ij,vg}^{l+1}\sum_{w}^{N_w}(\mathbf{B}_{i,gg,jw}^{l}\mathbf{A}_{iw,gv}^{l}\mathbf{v}_{iw}^{l-1}) + \mathbf{A}_{ij,vg}^{l+1}\sum_{c}^{N_c}(\mathbf{B}_{i,gG,jc}^{l}\mathbf{G}_{c,c}^{l}) + Cst \\
=& \mathbf{A}_{ij,vv}^{l+1}\mathbf{A}_{ij,vv}^{l}\mathbf{v}_{ij}^{l-1} + \sum_{w}^{N_w}(\mathbf{A}_{ij,vg}^{l+1}\mathbf{B}_{i,gg,jw}^{l}\mathbf{A}_{iw,gv}^{l}\mathbf{v}_{iw}^{l-1}) \\
&+ \sum_{c}^{N_g}(\sum_{w}^{N_w}(\mathbf{A}_{ij,vg}^{l+1}\mathbf{B}_{i,gG,jc}^{l}\mathbf{B}_{c,Gg,cw}^{l}\mathbf{A}_{cw,gv}^{l}\mathbf{v}_{cw}^{l-1})) + Cst
\end{aligned} \tag{4.6}$$

We can now observe the indirect attention coefficients that contribute to the formation of the visual tokens $\mathbf{v}_{ij}^{l+1}$. Let $\mathbf{C}_{ij,cw}^{l+1}$ represent the indirect attention of the visual tokens from crop i and window j with respect to all other crops c and windows w in the volume, for layer $l+1$. Consequently, we have the following expression:

$$\mathbf{C}_{ij,cw}^{l+1} = \begin{cases} \mathbf{A}_{ij,vg}^{l+1}\mathbf{B}_{i,gG,jc}^{l}\mathbf{B}_{c,Gg,cw}^{l}\mathbf{A}_{cw,gv}^{l} & \text{if } c \neq i \\ \mathbf{A}_{ij,vg}^{l+1}\mathbf{B}_{i,gG,ji}^{l}\mathbf{B}_{i,Gg,iw}^{l}\mathbf{A}_{iw,gv}^{l} + \mathbf{A}_{ij,vg}^{l+1}\mathbf{B}_{i,gg,jw}^{l}\mathbf{A}_{iw,gv}^{l} & \text{if } c = i \text{ and } w \neq j \\ \mathbf{A}_{ij,vg}^{l+1}\mathbf{B}_{i,gG,ji}^{l}\mathbf{B}_{i,Gg,iw}^{l}\mathbf{A}_{iw,gv}^{l} + \mathbf{A}_{ij,vg}^{l+1}\mathbf{B}_{i,gg,jj}^{l}\mathbf{A}_{ij,gv}^{l} + \mathbf{A}_{ij,vv}^{l+1}\mathbf{A}_{ij,vv}^{l} & \text{else.} \end{cases} \tag{4.7}$$

Fig. 4.1 illustrates the indirect attention computed using the formulation in equation Eq. 4.7. The visualizations demonstrate that LORI effectively captures long-range and out-of-range interactions. Notably, on the axial plane, we observe that the attention within the window (blue square) appears denser, indicating direct attention focused on this region. This observation aligns with the formulation in Eq. 4.7, where a higher number of attention weights are involved within the window. Moreover, the indirect attention outside the window exhibits diversity, with distinct focuses on other anatomical structures such as organs, bones, or skin boundaries. This indirect attention provides the model with robust positional information, enhancing its ability to accurately segment the pixel of interest.

4.5 Experiments

4.5.1 Datasets

In order to demonstrate the effectiveness of our proposed method, we conducted training and evaluation experiments on three distinct datasets: WORD [37], Synapse multi-organ segmentation (BCV) [38], and a private dataset named LIVUS. For each dataset, we report the results along with their corresponding standard deviations, which were computed across all different patients in the datasets.

WORD. The Whole abdominal ORgan Dataset (WORD) is a recently introduced large-scale dataset specifically designed for algorithm research and clinical application development. It consists of a collection of 150 abdominal CT volumes. In our experiments, we adhere to the predefined training and testing splits provided by the dataset, which encompass 100 volumes for training purposes and 30 volumes for testing. Each CT volume in the dataset is made up of a total of 16 distinct organs, namely the liver, spleen, left kidney, right kidney, stomach, gallbladder, esophagus, pancreas, duodenum, colon, intestine, adrenal gland, rectum, bladder, left femur, and right femur.

BCV. The BCV dataset consists of 30 abdominal CT scan cases. Following the commonly used data split [35], 18 cases are selected to form the training set, while the remaining 12 cases are kept for testing purposes. In our evaluation, we report the performance of the model on 13 specific abdominal organs, namely the spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava (IVC), portal vein system (PVS), pancreas, right adrenal gland, and left adrenal gland.

LIVUS. The LIver and Vessels UltraSound (LIVUS) dataset is a private segmentation dataset comprising 24 3D ultrasound volumes of the liver and its associated vessels: portal vein (PV), inferior vena cava (IVC), and hepatic vein (HV). The segmentation of these vessels is hard as they are similar and close one to another. For our experiments, we used a split of 16 volumes for training and 8 volumes for testing. The acquisition of this dataset was carried out by us. Three domain experts annotated each volume. The final segmentation masks are obtained by combining these annotations through the STAPLE[145] (Simultaneous Truth and Performance Level Estimation) algorithm. Four classes are annotated in this dataset: the liver, PV, IVC and HV.

4.5.2 Implementation details

This work was performed using HPC resources from GENCI-IDRIS, using a single Nvidia V100 GPU with 32GB memory for the experiments. Following the training strategy described in [34, 14, 35,

37], the initial learning rate is set to 0.01, and a polynomial decay strategy is employed for the learning rate schedule. The SGD optimizer was used, with a momentum value of 0.99 and a weight decay of 3e-5. Models are trained with deep supervision with a hybrid loss combining both cross-entropy and dice losses through addition. Training is performed for 1000 epochs, each epoch comprising 250 iterations with a batch size of 2.

As stated in [34], all volumes were first resampled to achieve a consistent target spacing. Then, a series of augmentations such as rotation, scaling, Gaussian noise, Gaussian blur, brightness and contrast adjustment, simulation of low resolution, and gamma augmentation were employed. However, we did not use mirroring augmentation due to its incompatibility with the proposed approach, which is strongly based on the positional information.

4.5.3 Backbones

In this study, we evaluate the efficiency of the proposed approach. To this end, we propose a comparison with various SOTA architectures: CNN with nnUNet[34] and DeepLabV3(2D) [39], Transformer with 3D Swin-UNet [35] and FINE [40], and hybrids with CoTr[14] and UNETR [36]. Except for DeepLabV3 and UNETR because of computing resources, all the models were trained in the same configuration, for comparison purposes.

LORI, being a versatile module, can be seamlessly integrated into various segmentation models. In this study, we chose nnUNet as the backbone for LORI due to its SOTA performance across multiple datasets. Additionally, nnUNet's convolutional-based architecture presents a limitation in terms of small receptive fields in high-resolution feature maps. However, this limitation is effectively addressed by LORI, which extends the model's ability to capture long-range interactions. To integrate LORI into nnUNet, we implemented a Swin Transformer module following the operations described in Sec. 4.4. This module was inserted after each convolutional layer of the nnUNet encoder, allowing LORI to utilize the feature maps generated by the convolutions. Additionally, in Sec. 4.6.3, we extended the application of LORI to other backbones such as 3D Swin-UNet. This was achieved by replacing the original operations with LORI's operations. Furthermore, in the case of CoTr, we incorporated extra global tokens into the sampling points set[124] and applied the out-of-range interaction operations on this modified set, enabling the integration of LORI into the CoTr model.

4.6 Results

4.6.1 Preliminary results

The results in this section are from preliminary experiments done on the BCV dataset with a smaller set of organs to match other prior methods papers showing results. Here we show that our proof of concept works by comparing it to multiple SOTA methods.

Method	Average		Per organ dice score (%)						
	HD95	DSC	Sp	Ki	Gb	Li	St	Ao	Pa
UNet [146]	-	77.4	86.7	73.2	69.7	93.4	75.6	89.1	54.0
AttUNet [19]	-	78.3	87.3	74.6	68.9	93.6	75.8	89.6	58.0
VNet [147]	-	67.4	80.6	78.9	51.9	87.8	57.0	75.3	40.0
Swin-UNet [21]	21.6	78.8	90.7	81.4	66.5	94.3	76.6	85.5	56.6
nnUNet [140]	10.5	87.0	91.9	86.9	**71.8**	**97.2**	85.3	**93.0**	**83.0**
TransUNet [107]	31.7	84.3	88.8	84.9	72.0	95.5	84.2	90.7	74.0
UNETR [36]	23.0	78.8	87.8	85.2	60.6	94.5	74.0	90.0	59.2
CoTr* [25]	11.1	85.7	93.4	86.7	66.8	96.6	83.0	92.6	80.6
nnFormer [2]	9.9	86.6	90.5	86.4	70.2	96.8	86.8	92.0	83.3
FINE*	**9.2**	**87.1**	**95.5**	**87.4**	66.5	97.0	**89.5**	91.3	82.5

Table 4.1: Method comparison using the BCV dataset and the training / test split from [2]. Average Dice scores are shown (DSC in % - higher is better). The average and individual organ 95% Hausdorff distances are also shown (HD95 in mm - lower is better). * denotes results trained by us using the authors' public code.

Single fold comparison To fairly compare with reported SOTA results, the same single split of 18 training and 12 test images was used as detailed in [2]. The results are provided in Table 4.1. FINE obtains the highest average Dice score of 87.1%, which is superior to all other baselines. It also attains the best average 95% Hausdorff distances (HD95) of 9.2mm. Note that the second best method in Dice (nnUNet) is largely below FINE in HD95 (10.5), and the second best method in HD95 (nnFormer) has a large drop in Dice (86.6).

5-fold cross-validation comparison A 5-fold cross-validation of 18 training and 12 test images was used to compare FINE with the public implementation of the leading transformer baselines (CoTr and nnFormer). The Dice score results are provided in Table 4.2. FINE's average improvement is significant (more than 1.5 pt with the second baseline with low variance), and FINE gives the best results in 6 out of 7 organ segmentation. The statistical significance in Dice is measured with a paired 2-tailed t-test. The significance of FINE gains with respect to CoTr (3e-2) and nnFormer (5e-2) is confirmed.

Method	Average	Sp	Ki	Gb	Li	St	Ao	Pa
CoTr [25]	84.4 ± 3.7	91.8 ± 5.0	87.9 ± 3.4	60.4 ± 10.0	95.7 ± 1.4	84.8 ± 1.3	**90.3** ± 1.8	80.0 ± 3.2
nnFormer [2]	84.6 ± 3.6	90.5 ± 6.1	87.9 ± 3.3	63.3 ± 8.1	95.7 ± 1.7	86.4 ± 0.8	89.1 ± 2.0	79.5 ± 3.5
FINE	**86.3** ± 3.0	**94.4** ± 1.9	**90.5** ± 4.3	**65.9** ± 7.8	**96.0** ± 1.1	**87.9** ± 1.2	89.4 ± 1.7	**80.2** ± 2.8
P-values	FINE vs. Cotr : 3e-2				FINE vs. nnFormer : 5e-2			

Table 4.2: Method comparison with SOTA transformer baselines (CoTr and nnFormer) using the BCV dataset and 5-fold cross validation. Results show mean and standard deviation of Dice (in %) for each organ and the average Dice over all organs (higher is better).

Method	WT	VT	Average		Per organ dice score						
			HD95	DSC	Sp	Ki	Gb	Li	St	Ao	Pa
nnFormer [2]	0	0	8.0	86.2	96.0	94.2	57.2	96.5	87.2	89.5	82.5
FINE	✓	0	7.7	86.6	95.7	94.2	60.9	**96.8**	85.1	90.0	**83.8**
	✓	✓	**5.2**	**87.1**	**96.2**	**94.5**	**61.5**	**96.8**	**87.3**	**90.3**	83.0

Table 4.3: Ablation study of the impact of different tokens on BCV dataset. The metrics are Dice score (DSC in %) for all organs and in average, and the 95% Hausdorff distance (HD95 in mm). WT: Window tokens. VT: Volume tokens.

Ablation study To show the impact of the different tokens in FINE, an ablation study is presented in Table 4.3. Three variations of FINE are compared: FINE without tokens, which is equivalent to the nnFormer method; FINE with window tokens but without volume tokens, and FINE with window and volume tokens (default). The results show that the window tokens generally help to better segment small and difficult organs like the pancreas (Pa) and gallbladder (Gb). The use of window tokens leads to an increase in average Dice by +0.4 points. Furthermore, adding volume tokens increases performance further (average Dice increase of +0.5 points, and average HD95 reduction from 7.7mm to 5.2mm).

4.6.2 Results

In this section, we conduct extended experiments on the three datasets presented earlier. Moreover, concerning the BCV dataset, we will now use all organs and retrain all SOTA methods as they do not always give the results for all classes in their original paper. By adding these difficult organs, we expect to show slightly lower average results. This setup is important to give fairer and reproducible results.

Comparisons with state-of-the-art. The experimental results, as presented in Tab. 4.4, highlight the superiority of our proposed method, LORI, when compared to state-of-the-art approaches across three diverse datasets. The evaluation demonstrates LORI's notable efficiency in accurately segmenting multiple organs within a 3D CT image. Specifically, LORI achieves a substantial dice score

improvement of +1.6 points on the WORD dataset and +0.35 points on the BCV dataset compared to the second best-performing method. Furthermore, LORI exhibits a commendable improvement of +0.86 points on the challenging LIVUS dataset, characterized by poor-quality ultrasound images. This outcome underscores LORI's ability to enhance the quality of segmentation in difficult modalities, further establishing its effectiveness in addressing challenging segmentation tasks.

Methods	**DeepLabV3+(2D)**	**UNETR(3D)**	**nnUNet**	**3D Swin-UNet**	**CoTr**	**FINE**	**LORI**
WORD	84.91	79.77	85.06	79.47	86.26	85.26	**87.86**
BCV	75.73	79.56	82.90	82.33	82.44	83.01	**83.25**
LIVUS	-	-	67.50	62.92	66.90	64.12	**68.36**

Table 4.4: Comparison of the overall organs mean Dice score (in %) of LORI with state-of-the-art methods on three different datasets: WORD, BCV, and LIVUS.

Methods	**DeepLabV3+(2D)**	**UNETR(3D)**	**nnUNet**	**3D Swin-UNet**	**CoTr**	**FINE**	**LORI**
Liver	96.21±1.34	94.67±1.92	96.57±0.63	96.36±0.57	96.54±0.60	96.36±0.66	**96.63±0.61**
Spleen	94.68±5.64	92.85±3.03	96.00±0.87	95.53±1.05	95.99±0.84	95.62±0.97	**96.10±0.89**
Kidney (L)	92.01±13.00	91.49±5.81	94.90±3.17	95.02±0.86	**95.70±0.83**	94.97±0.83	95.52±0.93
Kidney (R)	91.84±14.41	91.72±7.06	95.81±0.93	95.23±0.91	95.86±0.90	95.29±0.86	**95.90±0.87**
Stomach	91.16±3.07	85.56±6.12	91.88±2.85	91.23±2.78	**92.04±2.27**	91.00±3.17	91.94±2.76
Gallbladder	80.05±17.92	65.08±19.63	85.30±6.01	80.04±9.16	84.62±5.95	81.48±8.21	**85.69±5.97**
Esophagus	74.88±14.69	67.71±13.46	78.50±13.02	75.11±13.40	78.17±13.04	76.08±12.42	**78.86±12.09**
Pancreas	82.39±6.68	74.79±9.31	**85.46±5.59**	82.14±6.25	84.52±6.19	82.82±6.35	85.36±5.76
Duodenum	62.81±15.21	57.56±11.23	**70.38±15.34**	66.51±14.94	69.46±15.60	66.11±14.93	69.96±16.01
Colon	82.72±8.79	74.62±11.50	**87.35±8.68**	85.73±7.86	87.03±8.55	86.02±7.26	87.01±8.80
Intestine	85.96±4.02	80.40±4.59	**89.53±3.26**	88.15±3.13	89.52±2.97	88.36±2.94	89.36±3.20
Adrenal	66.82±10.81	60.76±8.32	**73.41±8.10**	66.21±9.74	71.55±8.78	66.96±9.81	72.76±8.50
Rectum	81.85±6.67	74.06±8.03	82.06±5.48	79.79±5.33	81.74±6.41	80.29±5.66	**82.35±4.88**
Bladder	90.86±14.07	85.42±18.17	92.28±9.91	91.61±10.38	91.94±10.75	91.37±11.04	**92.30±10.03**
Head of Femur (L)	92.01±4.76	89.47±6.40	76.62±16.64	54.60±6.70	82.02±20.44	91.46±4.57	**92.72±4.22**
Head of Femur (R)	92.29±4.01	90.17±4.00	64.82±14.46	28.32±6.54	83.38±21.93	79.99±12.23	**93.23±3.38**
Mean	84.91±5.05	79.77±4.92	85.06±3.12	79.47±3.05	86.26±3.95	85.26±3.03	**87.86±2.94**
P-values : LORI vs. nnUNet: 4.6e-8 ; LORI vs. 3D Swin-UNet: 1.7e-26 ; LORI vs. CoTr: 2.6e-3 ; LORI vs. FINE: 3.6e-03							

Table 4.5: Detailed results on WORD's dataset. The Dice score in % is given. The p-values between LORI and the methods we trained ourselves is also given.

In Tab. 4.5, we present the mean dice scores obtained for each organ on the WORD dataset, which offers the most diverse range of organs for segmentation. The results demonstrate that LORI outperforms competing methods on the majority of organs. Notably, LORI achieves significant improvements for organs such as the Gallbladder, Esophagus, and Femurs, which are less represented in the dataset in terms of organ distribution according to [37]. This suggests that LORI exhibits greater stability in segmenting organs with limited representation. A statistical t-test was conducted to evaluate the significance of the performance improvements achieved by LORI compared to nnUNet, 3D Swin-UNet, and CoTr. The p-values in Tab. 4.5, significantly below the significance level of 0.05,

provide strong evidence to confirm the statistical significance of the gains achieved by LORI compared to the other models.

Furthermore, as shown in Tab. 4.6, when considering the average 95% Hausdorff Distance, LORI achieves a value of 6.45 compared to 7.90 for the second-best method. The superior performance of LORI on this more stringent metric highlights its ability to accurately capture organ boundaries and produce more precise segmentations. The evaluation of the average symmetric surface distance (ASSD) further validates the superiority of LORI in terms of segmentation accuracy. With an ASSD of 0.97 compared to 1.86 for the second-best method, LORI demonstrates a substantial gain of 0.89. This outcome reinforces the notion that LORI outperforms other methods across multiple evaluation metrics.

Metric	**nnUNet**	**3D Swin-UNet**	**CoTr**	**LORI**
HD95	8.03 ± 4.56	30.10 ± 4.39	7.90 ± 5.41	$\mathbf{6.45 \pm 2.40}$
ASSD	1.86 ± 4.48	6.43 ± 1.35	1.87 ± 4.73	$\mathbf{0.97 \pm 0.61}$

Table 4.6: Results on WORD's dataset showing 95% Hausdorff distance (HD96) and average symmetric surface distance (ASSD) metrics.

Finally, LORI performs better on these datasets than FINE. Indeed, LORI is different from FINE for two main reasons: Firstly, LORI is a completely versatile module. This means that LORI can be used on any computer vision method to deal with out-of-range modeling problems. Secondly, LORI is more efficient than FINE as it only employs one transformer module to create the indirection link between all regions instead of two for FINE. This means that LORI needs fewer parameters to work and thus the training is also easier. These differences explained the better performances of LORI over FINE.

4.6.3 LORI Model analysis

Study of the versatility of LORI. To assess the individual contributions of long-range interaction and out-of-range interaction modeling in LORI, we conducted an ablation study on the WORD and LIVUS datasets, representing different modalities. By evaluating LORI with only local interaction, only long-range interaction, and the full LORI model which mixes long and out-of-range interaction, we were able to quantify the impact of each component. Additionally, we explored the versatility of LORI by employing different backbone architectures.

Tab. 4.7 presents the results of the ablation study, highlighting the influence of long-range interaction (GLAM) and its combination with out-of-range interaction (ORI) in LORI on the segmentation performance. Notably, utilizing GLAM methods alone consistently leads to improved average Dice

Method	Attention	#Params	LIVUS	WORD
nnUNet	local	31M	67.50±8.24	85.06±3.12
	GLAM	45M	68.03±8.84	86.15±3.83
	LORI	45M	**68.36±8.96**	**87.86±2.94**
3D Swin-UNet	local	159M	62.92±9.96	79.47±3.05
	GLAM	253M	60.19±9.74	79.88±2.97
	LORI	253M	**64.81±9.80**	**86.12±3.17**
CoTr	GLAM	41M	66.90±8.23	86.26±3.95
	LORI	63M	**67.19±8.59**	**87.48±3.08**

Table 4.7: Study of the importance of long-range mechanism (GLAM), the combination of long and out-of-range mechanisms (LORI) and versatility of LORI module. This ablation shows dice score in % on LIVUS and WORD dataset. The number of parameter of each method is given showing a small increase compared to the gains.

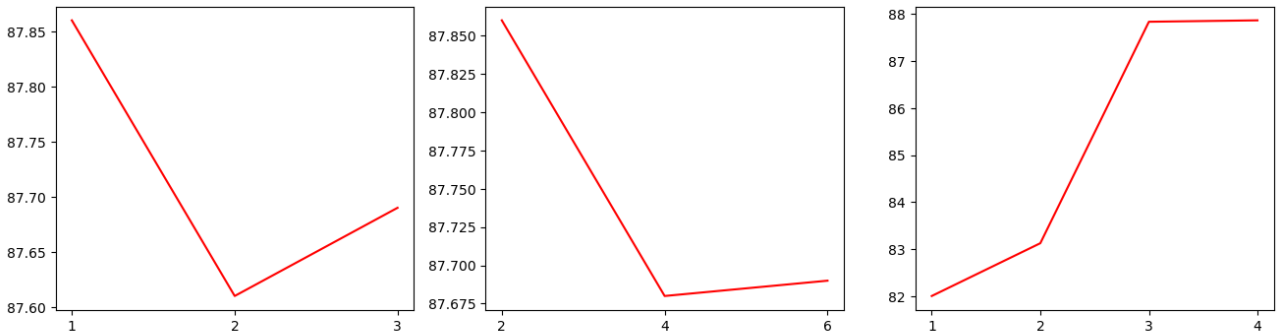
scores across the majority of cases. Furthermore, the incorporation of ORI methods consistently enhances the Dice scores, even in the presence of CoTr, which inherently integrates GLAM techniques.

The ablation study conducted in this research demonstrates the significance and complementarity of the two contributions introduced in LORI: long-range interaction and out-of-range interaction modeling. The results validate the importance of both components in improving segmentation performance. Moreover, the study highlights the potential of LORI as a valuable module to integrate into existing segmentation methods, as it has been shown to enhance overall performance.

Study of LORI's structure. The conducted ablation study, as illustrated in Fig. 4.5 (a) and (b), reveals that the addition of extra global tokens within the LORI or the addition of extra LORI modules is dispensable. This observation suggests that the improved performance achieved by LORI cannot be solely attributed to parameter augmentation, as an excessive increase in global tokens leads to a decline in performance. Additionally, Fig. 4.5 (c) demonstrates that the placement of LORI in the high-resolution layers yields superior performance compared to its exclusive placement in the low-resolution layers. This finding highlights the critical role of LORI in effectively modeling long-range interactions within high-resolution feature maps. It is important to note that the final architectural configuration of LORI adopted in this chapter was selected based on the best performance outcomes observed in these ablation studies.

Visualizations. Qualitative results, in the form of segmentation masks, are presented to offer a detailed analysis of LORI's performance in medical image segmentation. By visually comparing these results with those of other methods on the test set, the accuracy and effectiveness of LORI can be evaluated, providing valuable insights into its performance.

Fig. 4.6 showcases selected samples from the WORD dataset, providing visual evidence that



(a) Number of global tokens (b) Number of LORI module (c) Position of LORI

Figure 4.5: This ablation study shows the dice score (in %) on the WORD dataset of LORI with different architecture parameters. In (a) we experimented LORI with different number of global tokens associated with each region of the full-size image. In (b) we experimented with different number of LORI modules in a layer. In (c) we experimented different position of LORI in the based model: 1 correspond to only the last layer; 2 correspond to the two last layers; 3 correspond to the three last layers; 4 correspond to the four last layers.

reinforces our quantitative findings. LORI's ability to accurately segment less-represented organs is evident in the precise segmentation of the adrenal gland in the first row and the femur in the second row. Additionally, LORI demonstrates improved performance compared to 3D Swin-UNet in distinguishing between the right and left femur, thanks to its out-of-range interaction modeling that enhances spatial awareness. Furthermore, LORI outperforms other methods by avoiding the creation of holes in segmented organs, leveraging its utilization of the entire context outside the kidney which other models don't have access to because of the cropped patch size.

The LIVUS dataset serves as a valuable resource to assess the robustness of LORI, as depicted in Fig. 4.7. This dataset presents a challenging segmentation task, and LORI's strength is evident in its ability to address these difficulties effectively. Specifically, LORI demonstrates accurate segmentation by avoiding confusion between the inferior vena cava (IVC) and the portal vein (PV) in the first row of the visual results. Additionally, in the second row, LORI successfully captures the hepatic vein (HV) boundaries. Notably, LORI avoids segmenting the liver beyond its actual boundaries in both examples. These observations suggest that LORI leverages its long-range interaction modeling to accurately delineate complex organ boundaries such as IVC/HV, while the out-of-range interaction modeling aids in precisely identifying the location of the IVC and avoiding confusion with the PV. Moreover, the out-of-range interaction also ensures that regions of the image distant from the actual liver are not mis-segmented.

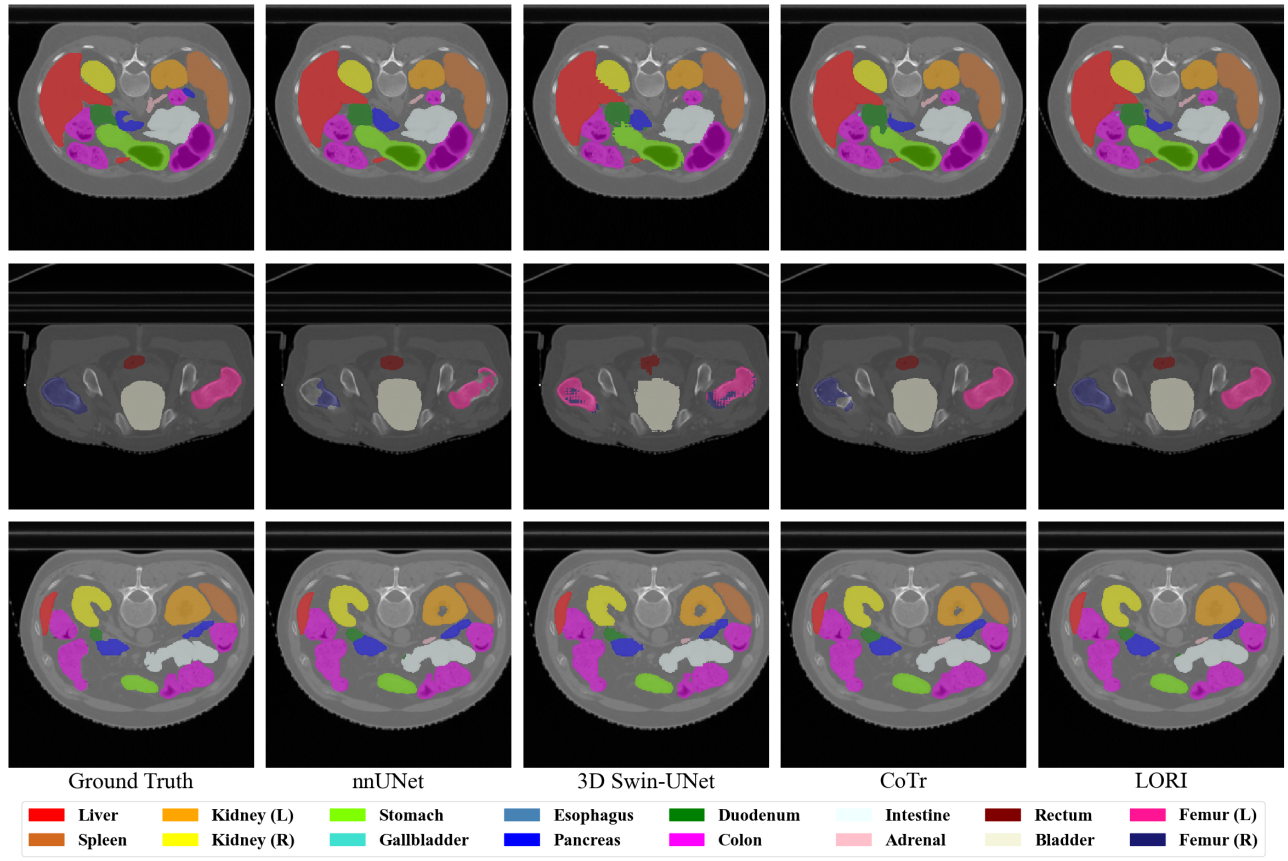


Figure 4.6: Visualisation of segmentation masks produced by the different methods on the test set of WORD dataset.

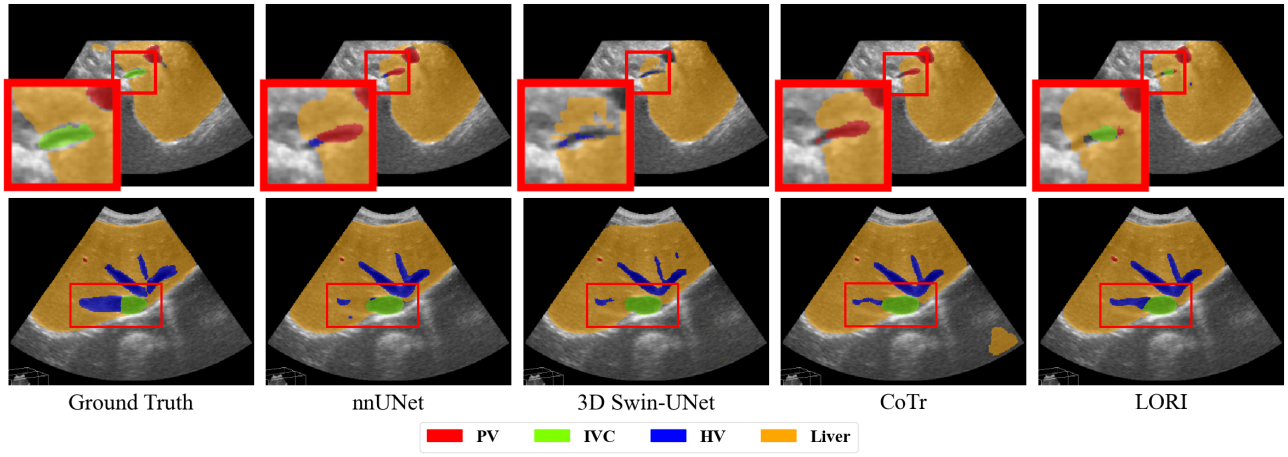


Figure 4.7: Visualisation of segmentation masks produced by the different methods on the test set of LIVUS dataset.

4.7 Conclusion

This chapter introduces LORI, a module specifically designed to address the challenges of long and out-of-range interaction modeling in 3D medical image segmentation. LORI enables information exchange among all visual tokens within a full-size volume, overcoming the limitations of small receptive fields in models such as nnUNet at high-resolution feature maps. By incorporating information from beyond the cropped patch, LORI effectively tackles a critical issue in 3D medical image segmentation. Experimental results demonstrate that LORI significantly improves segmentation performance on CT and ultrasound 3D image datasets. Importantly, LORI is a versatile module that can be seamlessly integrated into various existing models, consistently yielding performance enhancements.

Chapter 5

Conclusion and Perspectives

Contents

5.1	Contributions	**118**
5.2	On going work	**119**
5.3	Perspectives for futures Works	**119**

5.1 Contributions

In this thesis, the issue of segmenting high dimension images, specifically 3D medical images, was addressed. The high dimensionality of these images poses a challenge in their segmentation. It was explained that DL models require global context in order to effectively segment local regions, while classical models suffer from limited receptive fields or restricted input region sizes. To overcome these limitations, Transformers models were chosen for their ability to capture long-range interactions. We developed Transformers modules utilizing global tokens to enhance the capability of Transformers in modeling long and out-of-range interactions on high dimension images, such as 3D medical images.

U-Transformer. We first focused on the segmentation of 2D medical images and introduced a novel model called the U-Transformer. This model represents one of the earliest attempts to incorporate Transformer architectures for the purpose of medical imaging segmentation. Our contribution aimed to address the limitations of the classical UNet model, which has restricted receptive fields. To overcome this challenge, we integrated self-attention mechanisms in the encoder and multiple layers of cross-attention mechanisms in the decoder of our proposed U-Transformer model. By combining CNNs with Transformers, we developed a powerful hybrid model for the segmentation of 2D medical images. Our model achieved state-of-the-art performance on two different datasets and outperformed the nnUNet, which serves as a strong baseline in this domain.

Introduction of global tokens. The next step involved the development of GLAM, a module that can be seamlessly integrated into any windowed Transformer model. The primary objective of GLAM is to address the quadratic complexity of vanilla transformers but also of recent multi-resolution transformers, which are capable of handling high dimensional images but lack long-range interaction in high-resolution feature maps. GLAM leverages global tokens and specific Transformer modules to facilitate the propagation of information between windows. As a result, each window is interconnected, enabling the model to capture long-range information. The performance of GLAM surpassed that of traditional methods when evaluated on two real-life scene segmentation datasets and one 3D medical image segmentation dataset.

Long and out-of-range interaction modeling. Finally, we showed that the limited input size poses a significant challenge in the segmentation of 3D medical images. Traditional methods often employ smaller cropped patches to segment individual regions of the full volume, resulting in a substantial loss of contextual information. To address this issue, we propose LORI, which enables the modeling of long-and-out-of-range interactions and functions as a versatile module for various deep learning methods. Moreover, LORI efficiently utilizes global tokens by requiring only one transformer step to propagate information from the entire volume. We conducted experiments on three 3D medical images segmentation datasets, comprising two CT-scans and one ultrasound dataset. The results consistently

demonstrate that LORI outperforms classical methods across different configurations, highlighting its robustness and the significance of context modeling.

5.2 On going work

Medical image registration. Image registration consists of spatially aligning images representing identical structures, to overlap them and provide additional visual information. This process can be unimodal or multimodal, as well as intra-patient or interpatient. The integration of both segmentation and registration techniques is essential in developing a navigation system for percutaneous ultrasound-guided puncture. Furthermore, the application of segmentation and registration methods can complement each other in terms of performance and evaluation. For instance, surface-based registration necessitates a preliminary segmentation step, while transfer-based segmentation relies on registration. Additionally, the evaluation of registration methods often involves the utilization of segmentation masks, where the overlay after registration is evaluated.

Currently, we are collaborating with IRCAD on a public database comprising paired 3D kidney ultrasound and CT images, specifically designed for segmentation and registration purposes. As part of the evaluation process, the GLAM model has been employed as a baseline for segmentation. Fig. 5.1 shows some qualitative results obtained during this collaboration.

It would be also worth exploring with this new dataset the potential of our LORI method in the domain of multimodal image segmentation as well. By adapting global tokens, we could propagate information from one modality to another, thereby facilitating the segmentation of ultrasound images with the assistance of CT images.

5.3 Perspectives for futures Works

Medical videos. The challenge of dealing with medical US images has been described in this thesis. We have acknowledged the need for additional contextual information to enhance the performance of segmentation methods in accurately identifying anatomical structures within these images. One potential solution to incorporate more context is through the utilization of medical US videos. In contrast to static US images, medical US videos provide real-time data that can capture the entire process of acquiring a 3D ultrasound image. These videos contain valuable information and offer a richer context for analysis. However, the high-resolution of such videos poses a challenge for DL processing methods. To address this issue, the application of global tokens like GLAM or LORI presents an interesting solution. The use of global tokens enables efficient handling of high-resolution

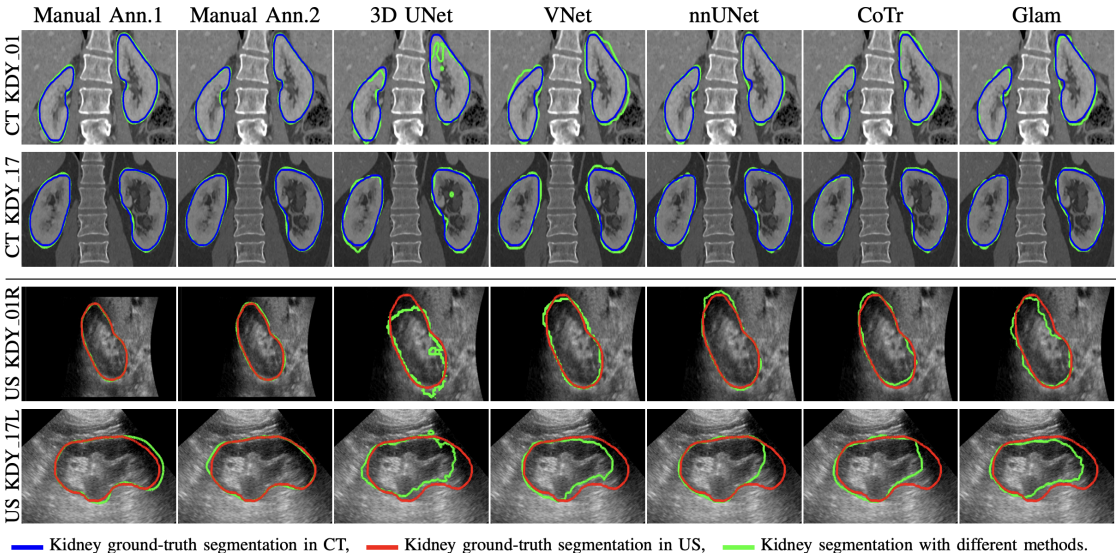


Figure 5.1: Qualitative results showing example CT and US segmentations using IRCAD's new dataset (patient 01 and 17). The top two rows shows coronal CT slices, with ground-truth segmentations overlaid in green, and estimated segmentations in blue. The bottom two rows shows longitudinal US slices, with ground truth segmentations overlaid in red, and estimated segmentations in blue. The two rows last (Annotator 1 and 2) show segmentations from each annotator, and the remaining rows show the best training version on average between single or double target(s) of each automatic segmentation from 5 DNN-based methods.

US videos, overcoming the limitations of traditional approaches. Furthermore, by utilizing global tokens, the treatment of the temporal aspect of the data could be enhanced, by allowing the propagation of information through time. This facilitates the modeling of long-range interactions over the duration of the video.

Real life experiments. IRCAD, through DISRUMPERE, aimed at getting tangible real-world impact through its research endeavors. We would like to further evaluate and also adapt LORI for real-time segmentation of US images. The envisioned outcome of this is the development of a sophisticated medical device capable of autonomously detecting tumors, delineating their boundaries, and subsequently executing an automatically precise surgical intervention via minimally invasive puncture procedures. It is important to underscore that this initiative represents a concerted interdisciplinary effort, involving multiple research departments, and holds the potential to signify a significant advancement within the domain of AI applied to the field of medicine.

Foundation models. Foundation models have been identified as a significant advancement in the field of AI for medical image analysis, as discussed in Moor's study [41]. These models combine text and images using a vast dataset, enabling them to perform multiple tasks without relying on task-specific labeled data. Additionally, the integration of Natural NLP components in these models provides

further insights into their decision-making process. Furthermore, the incorporation of global tokens in this research has the potential to enhance the models' ability to process high-dimensional data and address multimodal tasks effectively.

Global tokens in other tasks. The concept of global tokens is a versatile and adaptable tool that holds potential utility across diverse domains. In the realm of computer vision, for instance, the use of methods such as GLAM or LORI can be applied to address a spectrum of tasks involving high-dimensional image data. These tasks encompass but are not limited to satellite image analysis, 3D image reconstruction, image super-resolution, analysis of astronomical images, and video analysis. Furthermore, the versatility of global tokens transcends the confines of computer vision and extends to various other disciplinary domains. In the context of multi-task learning, global tokens may be leveraged to facilitate the propagation of information as necessary, enhancing the model's ability to perform multiple concurrent tasks effectively. Additionally, in instances where graph neural networks grapple with extensive graph structures, the integration of global tokens can enhance their scalability and information exchange capacities. Finally, within the domain of audio analysis, global tokens can be employed across various temporal ranges, enabling communication and synergy between them to bolster the analysis of complex audio data. Thus, the concept of global tokens as presented in this thesis manifests as a promising and cross-disciplinary approach with the potential to augment a wide spectrum of applications.

Bibliography

[1] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen et K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, n°. 2, p. 203–211, 2021.

[2] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang et Y. Yu, "nnformer: Interleaved transformer for volumetric segmentation," 2021.

[3] Z. Liu, P. Luo, X. Wang et X. Tang, "Deep learning face attributes in the wild," dans *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[4] H. Duan, Y. Zhao, K. Chen, D. Lin et B. Dai, "Revisiting skeleton-based action recognition," dans *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, p. 2959–2968.

[5] Pylessons, "Introduction to speech recognition with tensorflow." [En ligne]. Disponible: https://pylessons.com/speech-recognition

[6] N. Agarwal, "Machine translation." [En ligne]. Disponible: https://medium.com/@nupur94/machine-translation-715d1f460c07

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg et L. Fei-Fei, "Imagenet large scale visual recognition challenge," 2015.

[8] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, Y. Li, B. Zhang, Y. Liang, L. Zhou, X. Xu, X. Chu, X. Wei et X. Wei, "Yolov6: A single-stage object detection framework for industrial applications," 2022.

[9] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth et B. Schiele, "The cityscapes dataset for semantic urban scene understanding," 2016.

[10] A. Krizhevsky, I. Sutskever et G. E. Hinton, "Imagenet classification with deep convolutional neural networks," dans *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou et K. Weinberger, édit., vol. 25. Curran Associates, Inc., 2012. [En ligne]. Disponible: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[11] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox et O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," dans *MICCAI*, 2016, p. 424–432.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit et N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[13] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz et Y. Zhang, "Segment anything model for medical image analysis: an experimental study," 2023.

[14] Y. Xie, J. Zhang, C. Shen et Y. Xia, "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation," 2021.

[15] Y. Tay, M. Dehghani, D. Bahri et D. Metzler, "Efficient transformers: A survey," 2022.

[16] S. Wang, B. Z. Li, M. Khabsa, H. Fang et H. Ma, "Linformer: Self-attention with linear complexity," *arXiv e-prints*, p. arXiv–2006, 2020.

[17] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin et B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021.

[18] W. Luo, Y. Li, R. Urtasun et R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," 2017.

[19] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[20] O. Ronneberger, P. Fischer et T. Brox, "U-net: Convolutional networks for biomedical image segmentation," dans *MICCAI*, 2015, p. 234–241.

[21] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian et M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," 2021.

[22] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken et C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, p. 60–88, 2017. [En ligne]. Disponible: https://www.sciencedirect.com/science/article/pii/S1361841517301135

[23] H. B. Yedder, B. Cardoen et G. Hamarneh, "Deep learning for biomedical image reconstruction: a survey," *Artificial Intelligence Review*, vol. 54, n°. 1, p. 215–251, aug 2020. [En ligne]. Disponible: https://doi.org/10.1007%2Fs10462-020-09861-2

[24] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, B. van Ginneken, M. Bilello, P. Bilic, P. F. Christ, R. K. G. Do, M. J. Gollub, S. H. Heckers, H. Huisman, W. R. Jarnagin, M. K. McHugo, S. Napel, J. S. G. Pernicka, K. Rhode, C. Tobon-Gomez, E. Vorontsov, J. A. Meakin, S. Ourselin, M. Wiesenfarth, P. Arbeláez, B. Bae, S. Chen, L. Daza, J. Feng, B. He, F. Isensee, Y. Ji, F. Jia, I. Kim, K. Maier-Hein, D. Merhof, A. Pai, B. Park, M. Perslev, R. Rezaiifar, O. Rippel, I. Sarasua, W. Shen, J. Son, C. Wachinger, L. Wang, Y. Wang, Y. Xia, D. Xu, Z. Xu, Y. Zheng, A. L. Simpson, L. Maier-Hein et M. J. Cardoso, "The medical segmentation decathlon," *Nature Communications*, vol. 13, n°. 1, jul 2022. [En ligne]. Disponible: https://doi.org/10.1038%2Fs41467-022-30695-9

[25] Y. Xie, J. Zhang, C. Shen et Y. Xia, "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation," 2021.

[26] J. Devlin, M.-W. Chang, K. Lee et K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[27] F. Milletari, N. Navab et S. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," dans *2016 Fourth International Conference on 3D Vision (3DV)*, 2016, p. 565–571.

[28] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh et J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," dans *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2018, p. 3–11.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser et I. Polosukhin, "Attention is all you need," 2023.

[30] T. The Cancer Imaging Archive, "Pancreas-ct." [En ligne]. Disponible: https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT

[31] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso et A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *International Journal of Computer Vision*, vol. 127, n°. 3, p. 302–321, 2019.

[32] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth et B. Schiele, "The cityscapes dataset for semantic urban scene understanding," dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, p. 3213–3223.

[33] B. Landman, Z. Xu, I. Eugenio, Juan, M. Styner, T. Robin, Langerak et A. Klein, "Multi-atlas labeling beyond the cranial vault," *MICCAI*, 2015.

[34] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert et K. H. Maier-Hein, "nnu-net: Self-adapting framework for u-net-based medical image segmentation," 2018.

[35] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang et Y. Yu, "nnformer: Interleaved transformer for volumetric segmentation," 2022.

[36] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. Roth et D. Xu, "Unetr: Transformers for 3d medical image segmentation," 2021.

[37] X. Luo, W. Liao, J. Xiao, J. Chen, T. Song, X. Zhang, K. Li, D. N. Metaxas, G. Wang et S. Zhang, "WORD: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image," *Medical Image Analysis*, vol. 82, p. 102642, 2022.

[38] S. B. info@sagebase.org. Synapse | sage bionetworks. [En ligne]. Disponible: https://www.synapse.org

[39] L.-C. Chen, G. Papandreou, F. Schroff et H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017.

[40] L. Themyr, C. Rambour, N. Thome, T. Collins et A. Hostettler, "Memory transformers for full context and high-resolution 3d medical segmentation," dans *Machine Learning in Medical Imaging*, C. Lian, X. Cao, I. Rekik, X. Xu et Z. Cui, édit. Cham: Springer Nature Switzerland, 2022, p. 121–130.

[41] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol et P. Rajpurkar, "Foundation models for generalist medical artificial intelligence," vol. 616, n°. 7956, p. 259–265. [En ligne]. Disponible: https://doi.org/10.1038/s41586-023-05881-4

[42] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick et P. Dollár, "Microsoft coco: Common objects in context," 2015.

[43] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever et D. Amodei, "Language models are few-shot learners," 2020.

[44] OpenAI, "Gpt-4 technical report," 2023.

[45] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen et I. Sutskever, "Zero-shot text-to-image generation," 2021.

[46] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng et A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET Image Processing*, vol. 16, n°. 5, p. 1243–1267, jan 2022. [En ligne]. Disponible: https://doi.org/10.1049%2Fipr2.12419

[47] J. Moorthy et U. D. Gandhi, "A survey on medical image segmentation based on deep learning techniques," *Big Data and Cognitive Computing*, vol. 6, n°. 4, 2022. [En ligne]. Disponible: https://www.mdpi.com/2504-2289/6/4/117

[48] J. Noble et D. Boukerroui, "Ultrasound image segmentation: a survey," *IEEE Transactions on Medical Imaging*, vol. 25, n°. 8, p. 987–1010, 2006.

[49] M. H. Mozaffari et W. Lee, "3d ultrasound image segmentation: A survey," 2016.

[50] D. L. Pham, C. Xu et J. L. Prince, "Current methods in medical image segmentation," *Annual Review of Biomedical Engineering*, vol. 2, n°. 1, p. 315–337, 2000, pMID: 11701515. [En ligne]. Disponible: https://doi.org/10.1146/annurev.bioeng.2.1.315

[51] N. Sharma et L. Aggarwal, "Automated medical image segmentation techniques," *Journal of medical physics / Association of Medical Physicists of India*, vol. 35, p. 3–14, 04 2010.

[52] D. Withey et Z. Koles, "Three generations of medical image segmentation: Methods and available software," *Int J Bioelectromag*, vol. 9, 01 2007.

[53] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, n°. 1, p. 62–66, 1979.

[54] M. Sezgin et B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging*, vol. 13, p. 146–168, 01 2004.

[55] R. Pohle et K. Tönnies, "Segmentation of medical images using adaptive region growing," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 4322, 01 2002.

[56] M. Dabass, S. Vashisth et R. Vig, *Effectiveness of Region Growing Based Segmentation Technique for Various Medical Images - A Study*, 03 2018, p. 234–259.

[57] C. Jia-xin et L. Sen, "A medical image segmentation method based on watershed transform," 10 2005, p. 634– 638.

[58] W. E. Higgins et E. J. Ojard, "Interactive morphological watershed analysis for 3d medical images," *Computerized Medical Imaging and Graphics*, vol. 17, n°. 4, p. 387–395, 1993, 3D Advanced Image Processing in Medicine. [En ligne]. Disponible: https://www.sciencedirect.com/science/article/pii/089561119390033J

[59] T. Kohlberger, M. Sofka, J. Zhang, N. Birkbeck, J. Wetzl, J. Kaftan, J. Declerck et S. K. Zhou, "Automatic multi-organ segmentation using learning-based segmentation and level set optimization," vol. 14, 09 2011, p. 338–345.

[60] D. Cremers, O. Fluck, M. Rousson et S. Aharon, "A probabilistic level set formulation for interactive organ segmentation," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 6512, p. 30–, 03 2007.

[61] A. Wimmer, G. Soza et J. Hornegger, "A generic probabilistic active shape model for organ segmentation," dans *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009*, G.-Z. Yang, D. Hawkes, D. Rueckert, A. Noble et C. Taylor, édit. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, p. 26–33.

[62] T. Kohlberger, M. Uzunbas, C. Alvino, T. Kadir, D. Slosman et G. Funka-Lea, "Organ segmentation with level sets using local shape and appearance priors," vol. 12, 09 2009, p. 34–42.

[63] A. Klein, B. Mensh, S. Ghosh, J. Tourville et J. Hirsch, "Mindboggle: Automated brain labeling with multiple atlases," *BMC medical imaging*, vol. 5, p. 7, 11 2005.

[64] H. Park, P. Bland et C. Meyer, "Construction of an abdominal probabilistic atlas and its application in segmentation," *IEEE transactions on medical imaging*, vol. 22, p. 483–92, 05 2003.

[65] E. Schreibmann, D. Marcus et T. Fox, "Multiatlas segmentation of thoracic and abdominal anatomy with level set-based local search," *Journal of applied clinical medical physics / American College of Medical Physics*, vol. 15, p. 4468, 09 2014.

[66] R. Wolz, C. Chengwen, K. Misawa, M. Fujiwara, K. Mori et D. Rueckert, "Automated abdominal multi-organ segmentation with subject-specific atlas generation," *IEEE transactions on medical imaging*, vol. 32, 06 2013.

[67] J. Iglesias et M. Sabuncu, "Multi-atlas segmentation of biomedical images: A survey," *Medical image analysis*, vol. 24, 12 2014.

[68] H. Wang, J. Suh, S. Das, J. Pluta, C. Craige et P. Yushkevich, "Multi-atlas segmentation with joint label fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 06 2012.

[69] A. Neumann et C. Lorenz, "Statistical shape model based segmentation of medical images," *Computerized Medical Imaging and Graphics*, vol. 22, n°. 2, p. 133–143, 1998. [En ligne]. Disponible: https://www.sciencedirect.com/science/article/pii/S0895611198000159

[70] T. Okada, M. G. Linguraru, M. Hori, Y. Suzuki, R. Summers, N. Tomiyama et Y. Sato, "Multi-organ segmentation in abdominal ct images," *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, vol. 2012, p. 3986–9, 08 2012.

[71] M. Hammon, A. Cavallaro, M. Erdt, P. Dankerl, M. Kirschner, K. Drechsler, S. Wesarg, M. Uder et R. Janka, "Model-based pancreas segmentation in portal venous phase contrast-enhanced ct images," *Journal of digital imaging : the official journal of the Society for Computer Applications in Radiology*, vol. 26, 03 2013.

[72] Y. Lecun et Y. Bengio, "Convolutional networks for images, speech, and time-series," 01 1995.

[73] K. Simonyan et A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[74] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke et A. Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014. [En ligne]. Disponible: http://arxiv.org/abs/1409.4842

[75] K. He, X. Zhang, S. Ren et J. Sun, "Deep residual learning for image recognition," 2015.

[76] J. Long, E. Shelhamer et T. Darrell, "Fully convolutional networks for semantic segmentation," p. 3431–3440.

[77] E. Shelhamer, J. Long et T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, n°. 4, p. 640–651, 2017.

[78] H. Zhao, J. Shi, X. Qi, X. Wang et J. Jia, "Pyramid scene parsing network," dans *CVPR*, 2017.

[79] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff et H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," dans *ECCV*, 2018.

[80] T. Xiao, Y. Liu, B. Zhou, Y. Jiang et J. Sun, "Unified perceptual parsing for scene understanding," dans *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, p. 418–434.

[81] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov et L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," 2019.

[82] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu et B. Xiao, "Deep high-resolution representation learning for visual recognition," 2020.

[83] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell et S. Xie, "A convnet for the 2020s," 2022.

[84] A. Kirillov, R. Girshick, K. He et P. Dollár, "Panoptic feature pyramid networks," 2019.

[85] T. Xiao, Y. Liu, B. Zhou, Y. Jiang et J. Sun, "Unified perceptual parsing for scene understanding." [En ligne]. Disponible: http://arxiv.org/abs/1807.10221

[86] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn et A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, n°. 2, p. 303–338, juin 2010.

[87] H. Caesar, J. Uijlings et V. Ferrari, "Coco-stuff: Thing and stuff classes in context," 2018.

[88] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso et A. Torralba, "Scene parsing through ADE20k dataset," dans *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, p. 5122–5130. [En ligne]. Disponible: http://ieeexplore.ieee.org/document/8100027/

[89] Özgün Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox et O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," 2016.

[90] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser et I. Polosukhin, "Attention is all you need," 2017.

[91] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, mar 2020. [En ligne]. Disponible: https://doi.org/10.1016%2Fj.physd.2019.132306

[92] S. Hochreiter et J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, p. 1735–80, 12 1997.

[93] J. Devlin, M.-W. Chang, K. Lee et K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[94] S. Smith, M. Patwary, B. Norick, P. LeGresley, S. Rajbhandari, J. Casper, Z. Liu, S. Prabhumoye, G. Zerveas, V. Korthikanti, E. Zhang, R. Child, R. Y. Aminabadi, J. Bernauer, X. Song, M. Shoeybi, Y. He, M. Houston, S. Tiwary et B. Catanzaro, "Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model," 2022.

[95] X. Wang, R. Girshick, A. Gupta et K. He, "Non-local neural networks," 2018.

[96] S. Takase et N. Okazaki, "Positional encoding to control output sequence length," 2019.

[97] X. Chu, Z. Tian, B. Zhang, X. Wang et C. Shen, "Conditional positional encodings for vision transformers," 2023.

[98] G. Ke, D. He et T.-Y. Liu, "Rethinking positional encoding in language pre-training," 2021.

[99] A. P. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit et L. Beyer, "How to train your vit? data, augmentation, and regularization in vision transformers," *Transactions on Machine Learning Research*, 2022. [En ligne]. Disponible: https://openreview.net/forum?id=4nPswr1KcP

[100] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár et R. Girshick, "Segment anything," 2023.

[101] Y. Wang, D. Ni, H. Dou, X. Hu, L. Zhu, X. Yang, M. Xu, J. Qin, P.-A. Heng et T. Wang, "Deep attentive features for prostate segmentation in 3d transrectal ultrasound," *IEEE Transactions on Medical Imaging*, vol. 38, n°. 12, p. 2768–2778, dec 2019. [En ligne]. Disponible: https://doi.org/10.1109%2Ftmi.2019.2913184

[102] C. Li, Q. Tong, X. Liao, W. Si, Y. Sun, Q. Wang et P.-A. Heng, *Attention Based Hierarchical Aggregation Network for 3D Left Atrial Segmentation: 9th International Workshop, STACOM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers*, 02 2019, p. 255–264.

[103] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker et D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," 2019.

[104] D. Nie, Y. Gao et L. Wang, *ASDNet: Attention Based Semi-supervised Deep Networks for Medical Image Segmentation: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV*, 09 2018, p. 370–378.

[105] A. G. Roy, N. Navab et C. Wachinger, "Concurrent spatial and channel squeeze excitation in fully convolutional networks," 2018.

[106] A. Sinha et J. Dolz, "Multi-scale self-guided attention for medical image segmentation," 2020.

[107] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille et Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[108] R. Child, S. Gray, A. Radford et I. Sutskever, "Generating long sequences with sparse transformers," *arXiv preprint arXiv:1904.10509*, 2019.

[109] A. Roy, M. Saffar, A. Vaswani et D. Grangier, "Efficient content-based sparse attention with routing transformers," 2020.

[110] N. Kitaev, Łukasz Kaiser et A. Levskaya, "Reformer: The efficient transformer," 2020.

[111] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, D. Belanger, L. Colwell et A. Weller, "Rethinking attention with performers," 2022.

[112] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo et L. Shao, "Pvtv2: Improved baselines with pyramid vision transformer," *arXiv preprint arXiv:2106.13797*, 2021.

[113] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik et C. Feichtenhofer, "Multiscale vision transformers," *arXiv preprint arXiv:2104.11227*, 2021.

[114] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo et L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," 2021.

[115] J. W. Rae, A. Potapenko, S. M. Jayakumar, C. Hillier et T. P. Lillicrap, "Compressive transformers for long-range sequence modelling," dans *International Conference on Learning Representations*, 2019.

[116] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi et Y. W. Teh, "Set transformer: A framework for attention-based permutation-invariant neural networks," dans *Proceedings of the 36th International Conference on Machine Learning*, 2019, p. 3744–3753.

[117] I. Beltagy, M. E. Peters et A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[118] J. Ainslie, S. Ontañón, C. Alberti, V. Cvicek, Z. Fisher, P. Pham, A. Ravula, S. Sanghai, Q. Wang et L. Yang, "Etc: Encoding long and structured inputs in transformers," dans *EMNLP (1)*, 2020.

[119] P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang et J. Gao, "Multi-scale vision longformer: A new vision transformer for high-resolution image encoding," *ICCV 2021*, 2021.

[120] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu et V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," 2021.

[121] Y. Zhang, H. Liu et Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," 2021.

[122] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov et S. Zagoruyko, "End-to-end object detection with transformers," 2020.

[123] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr et L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," 2021.

[124] X. Zhu, W. Su, L. Lu, B. Li, X. Wang et J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," 2021.

[125] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit et N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[126] Z. Zhang, H. Zhang, L. Zhao, T. Chen et T. Pfister, "Aggregating nested transformers," dans *arXiv preprint arXiv:2105.12723*, 2021.

[127] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso et A. Torralba, "Semantic understanding of scenes through the ade20k dataset," 2018.

[128] Synapse, "Multi-atlas labeling beyond the cranial vault - workshop and challenge," *MICCAI*. [En ligne]. Disponible: https://www.synapse.org/#!Synapse:syn3193805/wiki/217789

[129] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng et S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," dans *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, p. 558–567.

[130] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu et Y. Wang, "Transformer in transformer," 2021.

[131] C.-F. Chen, Q. Fan et R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," 2021.

[132] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia et C. Shen, "Twins: Revisiting the design of spatial attention in vision transformers," 2021.

[133] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan et L. Zhang, "Cvt: Introducing convolutions to vision transformers," *arXiv preprint arXiv:2103.15808*, 2021.

[134] A. Buades, B. Coll et J.-M. Morel, "A non-local algorithm for image denoising," dans *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02*, ser. CVPR '05. Washington, DC, USA: IEEE Computer Society, 2005, p. 60–65. [En ligne]. Disponible: http://dx.doi.org/10.1109/CVPR.2005.38

[135] M. Contributors, "MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark," https://github.com/open-mmlab/mmsegmentation, 2020.

[136] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li et L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," dans *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, p. 248–255.

[137] F. Milletari, N. Navab et S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," dans *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, p. 565–571.

[138] O. Ronneberger, P. Fischer et T. Brox, "U-net: Convolutional networks for biomedical image segmentation," dans *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, p. 234–241.

[139] Y. Chang, H. Menghan, Z. Guangtao et Z. Xiao-Ping, "Transclaw u-net: Claw u-net with transformers for medical image segmentation," 2021.

[140] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen et K. Maier-Hein, “nnu-net: a self-configuring method for deep learning-based biomedical image segmentation.” *Nature methods*, 2020.

[141] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei et W. Liu, “Ccnet: Criss-cross attention for semantic segmentation,” 2019.

[142] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang et H. Lu, “Dual attention network for scene segmentation,” 2019.

[143] M. Yin, Z. Yao, Y. Cao, X. Li, Z. Zhang, S. Lin et H. Hu, “Disentangled non-local neural networks,” 2020.

[144] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr et L. Zhang, “Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers,” dans *CVPR*, 2021.

[145] H. B. Mitchell, *STAPLE: Simultaneous Truth and Performance Level Estimation*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, p. 233–236. [En ligne]. Disponible: https://doi.org/10.1007/978-3-642-11216-4_21

[146] O. Ronneberger, P. Fischer et T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” dans *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, p. 234–241.

[147] F. Milletari, N. Navab et S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” 2016.

[148] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser et I. Polosukhin, “Attention is all you need,” dans *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan et R. Garnett, édit., vol. 30. Curran Associates, Inc., 2017. [En ligne]. Disponible: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Appendix A

Detailed Non Local Upsampling

In our approach we introduce the Non-Local Upsampling (NLU) module which is used in place of conventional upsampling operations which are based on local information only (bilinear, deconvolution). As a contrary, the idea of the NLU is to upsample the semantic features based on all the tokens coming from the skip connection by using a MSA block in the Swin with Upernet or Swin-Unet heads.

The NLU module is detailed in Fig. A.1. By using the same blocks as in [148], the skip connection is embedded into a query matrix $\boldsymbol{Q} \in \mathbb{R}^{(4N_p)\times C}$ while the keys and values are computed from the semantic low resolution features: $\mathbf{K} \in \mathbb{R}^{N_p \times C}$ and $\mathbf{V} \in \mathbb{R}^{N_p \times C}$. The resulting attention matrix is $\boldsymbol{A} \in \mathbb{R}^{(4N_p)\times N_p}$. To maintain the residual connection in the Transformer block, the low resolution features are upsampled and a linear projection adapts the number of channels before the sum. Then a Feed Forward (FF) layer is also used. It is worth noting that a normalization layer is included in both parts but omitted in the schema for clarity. At the end, a concatenation of the skip-connection and the upsampled semantic features ends the NLU the same way than in the standard U-Net architectures.

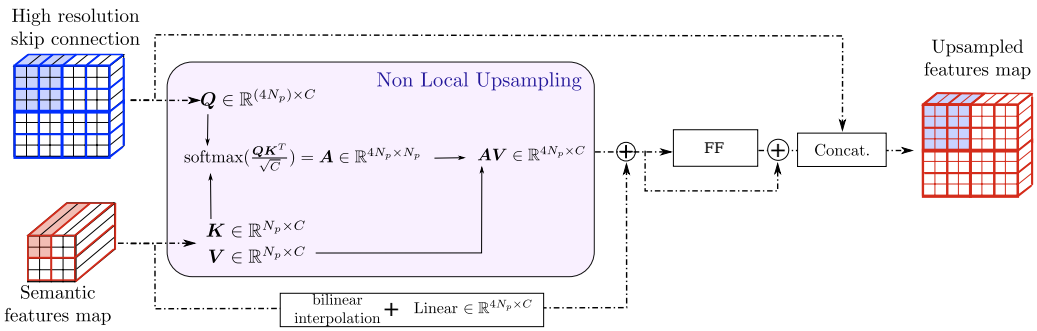


Figure A.1: **Non-Local Upsampling** The upsampling is processed window by window and is conceived as a super-resolution module where the low resolution feature map in the decoder (red) are re-embedded based on the high resolution ones coming from the encoder (blue). The patches are downsampled by a factor 2 before each hierarchy in the models. A given region from the decoder corresponds then to four neighbouring windows in the feature map coming from the skip connection.