Learning Image-to-Surface Correspondence

Riza Alp Guler

universite
PARIS-SACLAY

CentraleSupélec

NNT: 2019SACLC024

Apprentissage de Correspondances Image-Surface

Thèse de doctorat de l'Université Paris-Saclay
préparée à ***l'École CentraleSupélec***

École doctorale n° 580: Sciences et Technologies de l'Information et de la Communication (STIC)

Spécialité de doctorat ***Mathematiques & Informatique***

Thèse présentée et soutenue à Gif-sur-Yvette, le 8 Mars 2019 par

Rıza Alp Güler

Composition of the Jury :

Iasonas Kokkinos Associate Professor, University College London	Directeur de Thèse
Nikos Paragios Professor, École CentraleSupelec (CVN)	Co-encadrant
Ivan Laptev Research Director, INRIA Paris, & École normale supérieure	Rapporteur
Niloy Mitra Professor, University College London	Rapporteur
Andrea Vedaldi Associate Professor, University of Oxford	Examinateur
Stefanos Zafeiriou Associate Professor, Imperial College London	Examinateur
Isabelle Bloch Professor, Télécom ParisTech	Présidente

Thèse de doctorat

université
PARIS-SACLAY

CentraleSupélec

NNT: 2019SACLC024

Learning Image-to-Surface Correspondence

PHD Thesis to obtain the title of
Doctor of the Université Paris-Saclay

Doctoral School STIC (580)
Sciences et Technologies de l'Information et de la Communication
Speciality : ***Mathematiques & Informatique***

Thesis presented and defended at Gif-sur-Yvette on March 8, 2019 by

Rıza Alp Güler

Composition of the Jury :

Iasonas Kokkinos Associate Professor, University College London	**Advisor** [Directeur de Thèse]
Nikos Paragios Professor, École CentraleSupeléc (CVN)	**Co-Advisor** [Co-encadrant]
Ivan Laptev Research Director, INRIA Paris, & École normale supérieure	**Reviewer** [Rapporteur]
Niloy Mitra Professor, University College London	**Reviewer** [Rapporteur]
Andrea Vedaldi Associate Professor, University of Oxford	**Examiner** [Examinateur]
Stefanos Zafeiriou Associate Professor, Imperial College London	**Examiner** [Examinateur]
Isabelle Bloch Professor, Télécom ParisTech	**Chair** [Présidente]

Thèse de doctorat

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my thesis advisor Prof. Iasonas Kokkinos. His experience, accessibility and continuous interest in my work have made all the difference. I truly have learned a lot from him. He set an example as an advisor with his hands-on supervision style, dedication, fairness, scientific honesty and love of computer vision.

I am thankful to the thesis jury, Prof. Ivan Laptev, Prof. Niloy Mitra, Prof. Andrea Vedaldi, Prof. Stefanos Zafeiriou, and Prof. Isabelle Bloch for their valuable efforts in evaluating my thesis.

I am profoundly grateful to my collaborators for their contributions to the content of this thesis, Natalia Neverova, Siddhartha Chandra, Mihir Sahasrabudhe, Zhixin Shu, Stefan Kinauer, George Trigeorgis and others that are not mentioned. I also would like to thank Marie-Caroline for kindly helping with French translations. I was also fortunate to collaborate with Prof. Dimitris Samaras and Prof. Stefanos Zafeiriou.

I am grateful to Prof. Nikos Paragios, the founder of *Centre de Vision Numérique*, for everything that he has done for me and the lab. I will never forget the sincerity and camaraderie in CVN. I would like to thank my friends from CVN, Siddhartha, Mihir, Maria, Rafael, Marie-Caroline, Eugene, Stefan, Hari, Khue, Norbert, Evgenios and many more that are not mentioned for the wonderful years together.

I thank my parents Kerim and Nermin Güler for their unconditional love and support. I am grateful to my brother, H. Yiğit Güler, and his family for their moral support. I am also thankful to my brother for his help with the backend development of the annotation systems used in this thesis.

Finally, I thank my wife, Gizem, for providing me with motivation in life with her existence.

Apprentissage de Correspondances Image-Surface

Résumé

Cette thèse se concentre sur le développement de modèles de représentation dense d'objets 3-D á partir d'images. L'objectif de ce travail est d'améliorer les modèles surfaciques 3-D fournis par les systèmes de vision par ordinateur, en utilisant de nouveaux éléments tirés des images, plutôt que les annotations habituellement utilisées, ou que les modèles basés sur une division de l'objet en différents parties.

Des réseaux neuronaux convolutifs (CNNs) sont utilisés pour associer de manière dense les pixels d'une image avec les coordonnées 3-D d'un modèle de l'objet considéré. Cette méthode permet de résoudre très simplement une multitude de tâches de vision par ordinateur, telles que le transfert d'apparence, la localisation de repères ou la segmentation sémantique, en utilisant la correspondance entre une solution sur le modèle surfacique 3-D et l'image 2-D considérée. On démontre qu'une correspondance géométrique entre un modèle 3-D et une image peut être établie pour le visage et le corps humains.

Le chapitre 2 présente DenseReg, qui permet d'établir une correspondance dense entre les pixels d'une image et la représentation 3-D d'un visage. On propose d'utiliser un réseau neuronal convolutif qui permet de passer des coordonnées exprimées dans le domaine de l'image, á une paramétrisation continue et canonique du modèle 3-D. La méthode de la "régression quantifiée" est ensuite introduite, dans cette dernière on commence par sélectionner une position approximative quantifiée, qui est ensuite affinée grâce á la régression des résidus. Cette méthode permet d'établir l'état-de-l'art pour la localisation de repères sur un visage, ainsi que pour la segmentation de différentes parties d'un visage. L'approche proposée est également utilisée pour effectuer du "transfert de texture", en établissant une correspondance entre différentes instances de type objet.

Dans le chapitre 3, on démontre l'efficacité de la régression quantifiée pour l'estimation de pose humaine en volume 3-D. Les performances de la régression á propagation avant sont améliorées grâce á l'ajout d'une structure au modèle, qui impose des contraintes sur les positions relatives des différentes parties du corps. On utilise une technique d'inférence efficace basée sur le principe de séparation et d'évaluation, combinée á une inférence de modèles graphiques présentant différents niveaux de connectivité.

Le chapitre 4 introduit le principe de l'estimation dense de pose humaine, ou DensePose. Les problèmes de classification et de régression sont combinés pour établir une méthode qui permet de passer du domaine de l'image 2-D, á la paramétrisation continue de la surface du corps. On détaille une méthode efficace pour collecter des annotations de type image-vers-surface, qui sont développées spécifiquement pour le corps humain. Une base de donnée de grande échelle d'annotations, réalisées manuellement, est obtenue grâce á cette méthode. Une architecture CNN

basée sur une séparation en régions est ensuite présentée, cette dernière permet d'estimer de manière précise des correspondances pour chaque instance á une vitesse de plusieurs images par seconde.

Enfin, dans le chapitre 5, on utilise le principe de l'estimation dense de pose afin d'effectuer un transfert de pose humaine entre deux images. Ce problème revient á générer une nouvelle image d'une personne en se basant sur une unique image de cette personne, couplée á l'image d'une pose spécifique á transférer. L'efficacité de l'estimation dense de pose est montrée de manière quantitative pour le transfert de pose, par comparaison avec les techniques de division du corps en plusieurs parties, d'annotation et de segmentation.

Learning Image-to-Surface Correspondence

Abstract

This thesis addresses the task of establishing a dense correspondence between an image and a 3D object template. We aim to bring vision systems closer to a surface-based 3D understanding of objects by extracting information that is complementary to existing landmark- or part-based representations.

We use convolutional neural networks (CNNs) to densely associate pixels with intrinsic coordinates of 3D object templates. Through the established correspondences we effortlessly solve a multitude of visual tasks, such as appearance transfer, landmark localization and semantic segmentation by transferring solutions from the template to an image. We show that geometric correspondence between an image and a 3D model can be effectively inferred for both the human face and the human body.

We first propose dense shape regression, DenseReg, to establish dense correspondences between image pixels and a 3D face template. We propose a fully-convolutional neural network that maps coordinates from the image domain to a continuous, canonical parameterization of the template. We introduce 'quantized regression', a method that first selects a rough quantized position and then refines the localization through regression of the residuals. We report state-of-the-art performance in facial landmark localization and facial part segmentation tasks and also perform 'texture transfer' by establishing correspondences between different object instances.

We further demonstrate the effectiveness of quantized regression on volumetric 3D human pose estimation. We improve our feedforward regression results by adopting a structured model that imposes constraints between the relative positions of parts. We employ efficient inference using branch-and-bound and couple it with inference on graphical models with varying connectivity.

We then introduce the task of dense human pose estimation, or DensePose. We use a combination of classification and regression tasks to establish a mapping from the image domain to a continuous parametrization of the body surface. We propose an efficient pipeline for collecting image-to-surface annotations that is designed specifically for the human body and collect a large-scale manually annotated dataset. We then propose a region-based CNN architecture that regresses per-instance correspondences accurately at multiple frames per second.

We finally address the task of human pose transfer between two images by relying on the proposed dense pose estimation. This amounts to transferring the appearance of a person to a target pose. We quantitatively show the effectiveness of dense pose estimation for pose transfer by comparing to the alternatives of body parts, landmarks and segmentation masks.

Contents

List of Figures

CHAPTER 1

Introduction

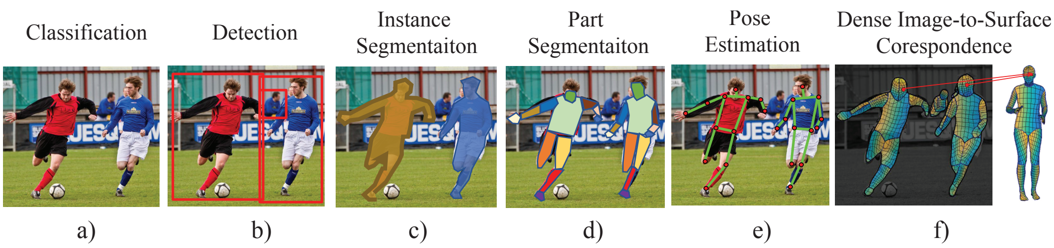


Figure 1.1: Progression of granularity in human understanding.

Understanding humans is at the core of current computer vision research due to its numerous applications such as human-computer interaction, augmented/virtual reality. The most basic form of human understanding would be the binary classification task, deciding about the presence of a person in the image (Fig. 1.1.a). Detection systems localize objects by producing boxes that contain persons (Fig. 1.1.b). Instance segmentation provides a more accurate localization by finding a mask for person instances in the image (Fig. 1.1.c). Part segmentation offers a more detailed understanding about the human, associating image regions with semantically meaningful body-parts (Fig. 1.1.d). What is currently commonly understood as 'human pose estimation' consists in localizing joints of the human body, reconstructing a skeleton as the human pose (Fig. 1.1.e). Rather than characterizing the region, or a few select points that relate to the object, in this thesis, we interpret an image through a mesh having thousands of nodes. We adopt a surface-based representation of the object of interest and establish correspondences between all foreground pixels on the image and a surface (Fig. 1.1.f).

The difference in the granularity of these tasks can be interpreted using the traditional divide of object understanding methods into (i) *discriminatively trained, bottom-up* and (ii) *deformable model-based* approaches.

Discriminative learning-based approaches, as those shown in Fig. 1.1.a-e, typically pursue invariance to shape deformations, for instance by employing local 'max-pooling' operations to elicit responses that are invariant to *local* translations. As such, these models can reliably detect patterns irrespective of their deformations through efficient, feedforward algorithms. At the same time, however, this discards useful shape-related information. Several recent works in deep learning have aimed at enriching deep networks with information about shape by explicitly modelling *the effect* deformations; having found success in classification [Papandreou 2015], fine-grained recognition [Jaderberg 2015], and also face detection [Chen 2016b]. In these works, the shape is treated as a nuisance, while we treat it as the goal in itself.

By contrast, approaches that rely on Statistical Deformabe Models (SDMs), such as Active Appearance Models [Cootes 2001] or 3D Morphable Models [Blanz 1999] aim at explicitly recovering dense correspondences between a deformation-free template and the observed image. SDM-based methods are limited in several respects. Firstly they require initialization from external systems, which can become increasingly challenging for elaborate SDMs. Furthermore, SDM fitting requires iterative, time-demanding optimization algorithms, especially when the initialization is far from the solution. Finally, the modelling and generalization capabilities of SDMs are bounded by the diversity of the dataset they are trained with.

Motivated by the gap between discriminatively trained systems for detection and category-level deformable models, we propose a framework that combines the merits of both. We declare correspondences from the image domain to a 2-dimensional, deformation-free parameterization of the template surface by training neural networks that densely regress the parameterized coordinates. This combines the fine-grained discriminative power of statistical deformable models with the "in the wild" operation of convolutional neural networks. The established correspondences are not necessarily bounded by the expressive power of a statistical model.

In the multi-object setting, the proposed task involves several other problems such as object detection, pose estimation, part and instance segmentation either as special cases or prerequisites. Addressing this task has applications in problems that require going beyond plain landmark localization, such as graphics, augmented reality, or human-computer interaction, and could also be a stepping stone towards general 3D-based object understanding. For the human body, existing 3D ground truth datasets such as the Human 3.6m dataset [Ionescu 2014b] does not carry information about the surface of the body. For instance, it is impossible to infer how fat a person is from a side pose given only joint annotations. On the other hand, the proposed dense correspondences provide information regarding the whole visible surface on the image.

In this chapter, we firstly list the contributions of the thesis in Sec. 1.1. To position our contributions within the broad range of works in the field of computer vision, we continue with the review of prior works in Sec. 1.2. We describe the structure of the thesis in Sec. 1.3 and list the publications and dissemination activities in Sec. 1.4.

1.1 Contributions of the thesis

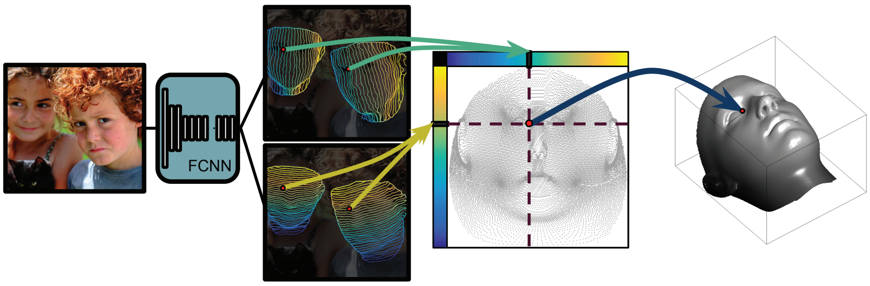


Figure 1.2: We propose *dense shape regression* for establishing correspondences between an image and an object template. Image pixels are localized on the surface by regressing a continuous, canonical parameterization of the template. Regressed correspondences are demonstrated for a point on the face template.

Surface alignment as regression via deep neural networks. We introduce the task of *dense shape regression* from RGB images. We parameterize the template shape in a two-dimensional deformation-free space, as visualized in Fig. 1.2. By regressing the location of points in this canonical space, we localize each foreground pixel on the template surface. We show that this regression problem can be solved accurately and efficiently using fully-convolutional neural networks and discriminative training.

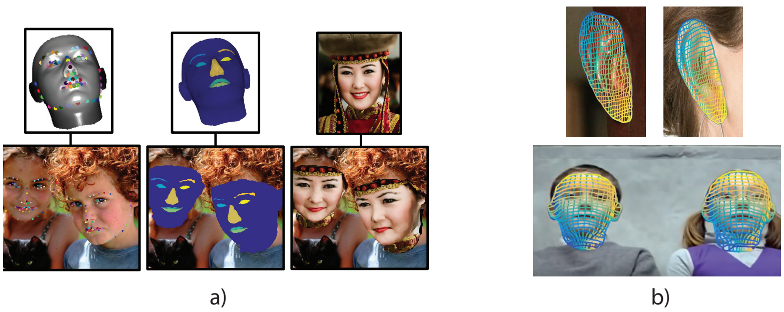


Figure 1.3: a) We solve image-level problems by backward-warping a canonical solution from the template coordinates to the image domain. Results for landmark localization, semantic part segmentation, and face transfer are demonstrated respectively. b) Estimated dense correspondences for the human ear and the human face.

We define a host of problems geometrically on the template domain, such as landmark localization, semantic segmentation and texture transfer. We solve such problems by transferring a fixed solution from template coordinates to the image using the estimated correspondences, as visualized in Fig. 1.3.a. We demonstrate the generic nature of the proposed method with applications on the human face and the human ear as depicted in Fig. 1.2.b. We report state-of-the-art quantitative results on facial landmark localization and facial part segmentation.

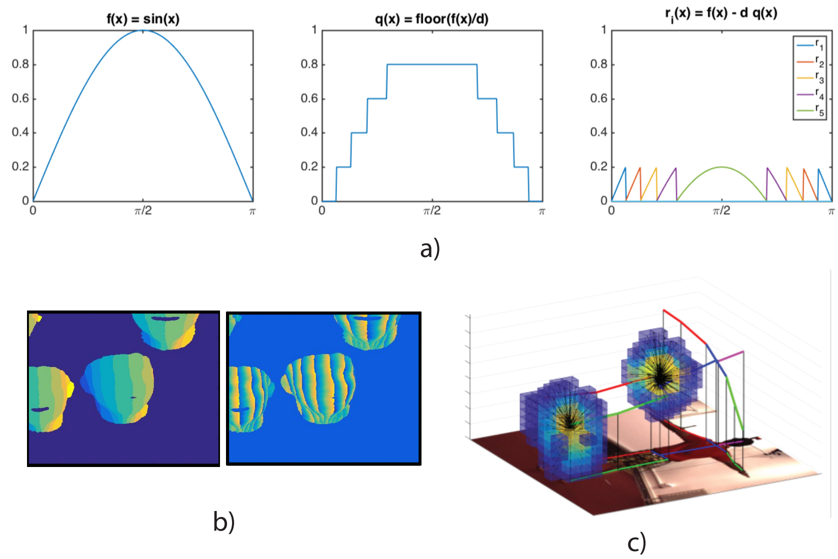


Figure 1.4: We propose the *quantized regression* algorithm, where the quantized signal is estimated by classification and residuals are regressed by separate regressors. a) A toy example showing the proposed separation for a sine wave. b) Classified quantized values and regressed residuals for the deformation-free coordinates of the human face. c) Quantized regression for monocular 3D human pose estimation.

Quantized Regression We draw inspiration from recent successes of object detection at the task of bounding box regression [Ren 2015] and introduce a method that blends classification and regression to accurately regress the template coordinates. The method involves selection of a rough quantized position and the regression of the residuals for better localization. We call this the 'quantized regression'. We estimate the quantized values through a classification branch and the residuals through regression units dedicated for each quantized value. The quantizated signal and residuals for each quantized value for a sine function are demonstrated in Fig. 1.4.a. We show that quantized regression outperforms naive regression of canonical coordinates and the granular classification of discretized coordinates. Estimated quantized coordinates and residuals for the human face are depicted in Fig. 1.4.b.

Monocular 3D Pose Estimation with Quantized Regression and Structured Prediction We show that the quantized regression strategy performs well for the localization of human joints volumetrically in monocular 3d human pose estimation. Instead of exclusively relying on a feed-forward architecture, we improve our estimation with a structured prediction algorithm that imposes constraints between the relative positions of parts. The quantized regression for localization of 3D human landmarks is visualized in Fig. 1.4.c, where a high resolution in pose estimation is achieved without increasing the computation/memory requirements.

Figure 1.5: We propose a system for *dense human pose estimation*, finding correspondences between human pixels and a 3D template of the human body.

Dense Human Pose Estimation Having demonstrated the feasibility of dense image-to-surface alignment for the face, we then turn to the substantially more challenging task of establishing correspondences between images and a 3D template of the human body, DensePose. We regress correspondences through local coordinate systems that we define for parts of the human body. The local coordinate systems and results of the DensePose system are depicted in Fig. 1.5.

To train the DensePose system, we have collected a large dataset of manually annotated correspondences using an efficient annotation pipeline. We use a region-based architecture that delivers per-instance dense correspondence results multiple frames per second.

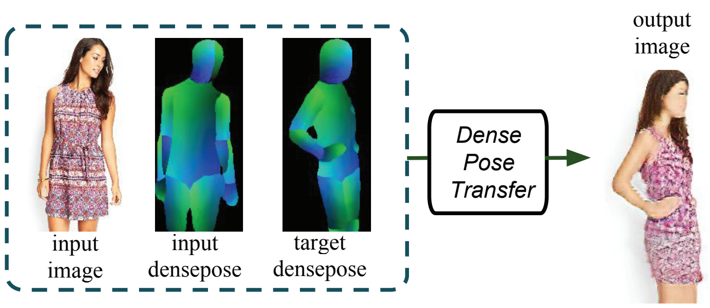


Figure 1.6: We introduce *dense pose transfer* for synthesis of a new image based on appearance and pose sourced from different input images.

DensePose radically improves the granularity of human understanding from images as demonstrated in Fig. 1.5, allowing geometric detail-demanding applications such as texture transfer for the human body for the first time.

Dense Pose Transfer Building on top of the DensePose system, we propose ‘dense pose transfer’ for transferring the appearance of a person to a target pose as demonstrated in Fig. 1.6. We integrate surface-based modeling with neural synthesis and fuse (i) a data-driven predictive model and (ii) a surface-based model that directly transfers the coordinates based on the dense correspondences. We account for occlusions by introducing an inpainting network that operates in the surface coordinate system. We quantitatively show the effectiveness of dense pose estimation for pose transfer by comparing to the alternatives of body parts, landmarks and segmentation masks.

1.2 Prior Work

In this section, we review the literature relevant to the contributions of the thesis. We start by introducing deep learning-based bottom-up techniques in Sec. 1.2.1 with a specific focus on the tasks involved in the thesis, such as object detection, instance segmentation, 2D and 3D human pose estimation. We then provide a review of deformable templates in Sec. 1.2.2 with a special focus on the 3D Morphable Models (3DMMs) for the human face and the body.

1.2.1 Discriminatively Trained, Bottom-Up Techniques

Bottom-up approaches for computer vision have been relying on local visual descriptors such as SIFT [Lowe 2004, Mikolajczyk 2005]. Handcrafted features have found broad use in solving problems, such as object detection [Dalal 2005], semantic segmentation [Shotton 2008] human pose estimation [Agarwal 2006a]. The features are typically blockwise orientation histograms, similar in function with the complex cells in V1, the first stage in the visual pathway of primates. They encode low-level perceptual information, whereas recognition requires higher-level visual processing. With the advent of deep learning [LeCun 1998, Krizhevsky 2012, Simonyan 2014b], the downstream Convolutional Neural Network (CNN) based features had led to a significant performance boost in recognition tasks in the field of computer vision. In this section, we introduce some of the accurate and robust techniques in detail for various problems. We focus on problems that are especially related to the contributions of this thesis.

1.2.1.1 Object Detection

Object detection is the process of localizing each object instance in an image and determining the class of each object. In computer vision the localization is typically done at a bounding box level.

The localization aspect of detection can be seen as a search problem. One typical search approach is to use a sliding window, with the basic assumption that the object can be located at any position and scale in the image. This exhaustive search was used in the first CNN based detection systems for faces [Vaillant 1994, Rowley 1998] and followed for pedestrians [Sermanet 2013b]. This is also common practice in detectors based on hand-crafted features, e.g. [Viola 2001, Dalal 2005, Harzallah 2009]. Alternatively, [Lampert 2009] shows that the search space can be reduced by exploiting the regular grid. This is done by a branch and bound technique operating with bounds provided by a linear classifier. Another alternative is to resort to class-agnostic region proposals obtained via grouping strategies. A popular example of such systems would be the 'selective search' [Uijlings 2013], which diversifies the search by proposing a variety of complementary image partitionings via hierarchical grouping. More recently [Ren 2015] proposes learning localization by classifying 'objectness' of fixed anchors on the image.

Deformable Part Models [Felzenszwalb 2008] revisited the idea of pictorial structures [Fischler 1973], and proposed discriminatively trained DPMs for object detection. DPMs had led to a significant performance improvement over existing baselines. However such modelling efforts were overshadowed by the bottom-up approaches when the hand-crafted features are replaced by CNN features [Sermanet 2013a, Girshick 2014].

Bottom-up systems such as [Dalal 2005], typically compute features, score every subwindow using a discriminatively trained classifier and finally apply non-maxima suppression to detect objects. The features, in this specific case HOG, encode low-level information about the objects, which can be constraining for recognition, especially when a shallow classifier is used.

The region-based CNN (R-CNN) of [Girshick 2014] crops images within selective search proposal boxes and extracts CNN features, which are then classified with an SVM. Fast-RCNN [Girshick 2015] pools features that correspond to regions of interest instead of cropping images, leading to significant speed improvements. Features pooled from a region go through fully connected layers that output class probabilities and bounding box regressions. This system is further improved in Faster-RCNN [Ren 2015], where a Region Proposal Network (RPN) replaces the selective search proposals. There are many more variants such as R-FCN [Dai 2016b] that is fully convolutional until the very end layer, where pooling takes place. [Lin 2017] proposes high-level semantic feature maps at smaller scales via lateral connections, which improves detection accuracy. Single shot systems such as SSD [Liu 2016b] and YOLO [Redmon 2016] directly classify anchor boxes and are typically faster.

1.2.1.2 Instance Segmentation

Segmentation is the process of dividing the image into regions that are meaningful for the 'purpose at hand' [Marr 1982]. The purpose can require the segmentation of semantic or functional regions, or correspondences to physical objects or their parts. The problem of semantic segmentation was typically approached by per-pixel classification of densely extracted features [He 2004, Shotton 2008]. These systems suffered from the lack of expressiveness of the features. The necessary context was not captured, and the individual per-pixel predictions were noisy. Earlier systems adopted conditional random fields (CRFs) that enforce similar labels for pixels that are close in appearance and spatial distance. Using CNNs for the task of dense pixel labeling led to significant improvements in terms of performance. A fully-convolutional architecture is introduced for dense labeling in the seminal work of [Long 2015]. [Chen 2018b] shows that convolution with upsampled filters, or ‘atrous convolution’ [Holschneider 1990] further improves performance.

Instance Segmentation requires both object detection and the foreground segmentation of the detected object instance. Methods for instance segmentation can be roughly divided into two categories, systems starting with the detection of the object and systems starting with the segmentation of the whole image.

Detection-first instance segmentation systems start by localizing objects or object candidates in the image. SDS [Hariharan 2014] and CFM [Dai 2015] propose systems where proposal regions are taken as input and refined through CNNs. In hypercolumns, [Hariharan 2015] exploits features from the intermediate regions for figure-ground segmentations starting from cropped images inside bounding box detections. The DeepMask and SharpMask systems [Pinheiro 2015, Pinheiro 2016] learn to propose candidate region segmentations and classify them. Similarly, [Dai 2016a] proposes a cascaded system where segmentation proposals are predicted and later classified. Similar to R-FCN, [Li 2017] predicts fully convolutional maps of object classes and foreground/background maps, allowing inference of instance segmentation masks. Mask-RCNN [He 2017] builds on top of the Faster-RCNN system [Ren 2015], adding a new branch that predicts the foreground mask, parallel to the bounding box recognition branch. [He 2017] also proposes the RoIAlign layer, which better respects the spatial locations of the features pooled. In Fig. 1.7, the architecture and qualitative results of the Mask-RCNN system on the COCO-dataset test set [Lin 2014] are depicted. More recently, MaskLab [Chen 2018a] builds on top of Faster-RCNN, fusing estimates of semantic segmentation and direction towards object center within each box to infer instance segmentations.

Segmentation-first instance segmentation systems typically depend on dense labelling. [Liang 2015] introduces the proposal free network for instance segmentation by densely predicting instance numbers along with category-level confidences and uses spectral clustering. [Zhang 2015b] proposes estimating the depth ordering of instances of objects to solve instance segmentation. [Uhrig 2016] densely predicts semantics, depth, and instance center direction. The predictions are used to compute template matching scores, which are fused to obtain instance segmentations. Deep Watershed Transform, [Bai 2017], predicts unit vectors pointing away from the nearest boundary and the distance transform for the objects to infer instances. [Liu 2017] predicts horizontal and vertical object breakpoints and sequentially composes object instances. In InstanceCut, [Kirillov 2017] exploits edges to infer instance segmentations. [De Brabandere 2017, Fathi 2017, Newell 2017] propose learning pixel-level embeddings, which are grouped to form instance segmentations. [Papandreou 2018] proposes a person instance segmentation system, where an embedding distance metric is defined based on estimated human keypoint locations.

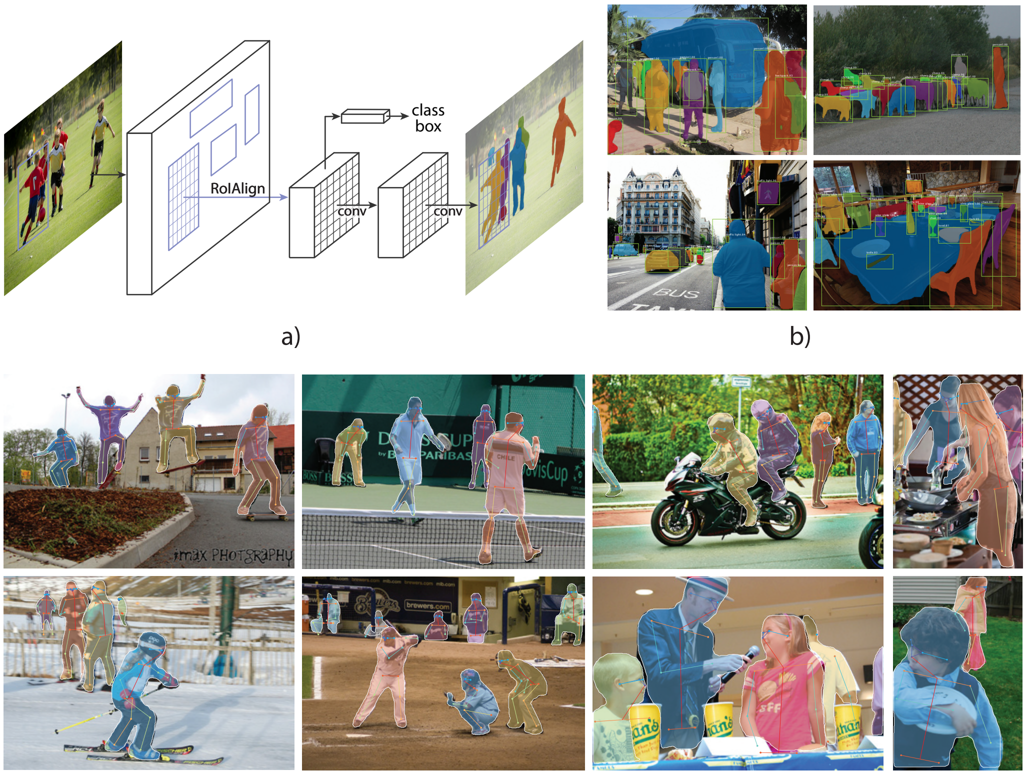


Figure 1.7: Demonstration of the Mask-RCNN system. a) The Mask-RCNN architecture: Task-specific fully convolutional networks operating on features pooled from region of interest. b) Mask-RCNN results for multi-class instance segmentation. c) Mask-RCNN results for human pose estimation and person instance segmentation. *Figures taken from the paper cited in the caption.*

1.2.1.3 Human Pose Estimation

What is conventionally referred to as human pose estimation is the problem of localizing anatomical keypoints defined on the human body, such as hips, elbows, ankles, etc.

Classical human pose estimation systems typically use graphical models that model the spatial dependencies between parts. Pictorial Structures [Felzenszwalb 2005] proposed a tree-structured graphical model that uses binary masks obtained via background subtraction. Following works use more sophisticated features with similar models [Andriluka 2009, Eichner 2009, Sapp 2010a, Johnson 2011, Dantone 2013, Sapp 2013, Pishchulin 2013, Yang 2013].

Deep learning based human pose estimation systems has drastically improved the performance of human pose estimation systems. [Tompson 2014, Chen 2014] propose systems that combine the graphical models with convolutional networks. In contrast, DeepPose [Toshev 2014] is a cascaded system, where spatial coordinates of keypoints are directly regressed from the image. A common practice is to employ cascaded or iterative estimation of the pose. [Carreira 2016] proposes 'Iterative Error Feedback'. [Wei 2016b] proposes convolutional pose machines (CPM) based on the previous multi-stage pose machines framework [Ramakrishna 2014]. The first stage takes the image as input and outputs localization heatmaps. The second stage takes both the image and estimated heatmaps as input and outputs the refined heatmaps. The second stage can be iteratively applied, refining the estimated localization heatmaps. [Belagiannis 2017] proposes a similar system with weight sharing, obtaining a recurrent system. [Newell 2016] proposes the stacked hourglass architecture, a fully-convolutional architecture with skip connections. Similar to previous work, they show the benefits of intermediate supervisions. Recently, [Yang 2017] reports further improvements with feature pyramids in the same 'hourglass' framework.

Multi-person pose estimation, just like the instance segmentation problem, is coupled with the detection of person instances. There are two common strategies.

Top-down approaches first detect the person instances, then infer the pose for each detected person post hoc. This allows methods for single-person pose estimation to be directly applied in the multi-person scenario, e.g. [Pishchulin 2012]. Many recent approaches that use deep learning adopt this approach effectively, e.g. G-RMI [Papandreou 2017], RMPE [Fang 2017], CPN [Chen 2017c]. A recent example is [Xiao 2018], where the authors propose a simple and quite effective baseline with several deconvolution operations on top of a standard fully-convolutional network operating on cropped images. Within the Mask R-CNN [He 2017] framework, as described in Sec. 1.2.1.2, keypoint localization can be implemented as another head, sharing the feature representation with the other tasks. Results obtained from this system is visualized in Fig. 1.7.

Bottom-up multi-person systems localize keypoints and then group them to infer human instances. [Pishchulin 2016, Insafutdinov 2016, Iqbal 2016] localize parts and perform grouping via integer linear programming. [Cao 2016] estimates not only heatmaps for localization but also direction fields between a keypoint and its parent. These direction fields, called 'part affinity fields', are utilized in declaring person instances. [Newell 2017] proposes learning dense embeddings to infer group instances. PersonLab [Papandreou 2018] proposes grouping using an embedding distance metric based on estimated offsets for keypoints. [Kocabas 2018] proposes a system that assigns keypoints to detected person instances.

1.2.1.4 Monocular 3D Pose Estimation

Monocular 3D pose estimation deals with 3D localization of relevant human keypoints given a single frame or video.

Estimation of 3D motion and pose for humans from videos has been a topic studied for more than three decades [O'rourke 1980]. Due to the lack of publicly available datasets, evaluation of early systems has been solely qualitative [Mori 2002, Brand 1999]. Some following works used synthetically generated data, e.g. [Shakhnarovich 2003, Grauman 2003, Sminchisescu 2005, Agarwal 2006b], yet the lack of photorealism of the rendered images makes the generalization to natural images problematic. [Sigal 2010] presented HumanEva, a publicly available dataset of synchronized motion capture (mocap) and multi-view video. Such ground truth data allows discriminative training of 3D localization systems, also allowing a fair evaluation of the performance of different approaches. The readers interested in methods previous to the availability of mocap based ground truth are referred to [Sigal 2010] for a chronological review. More recent datasets that provide mocap based ground truth are [Ionescu 2014b] and [Mehta 2017].

3D pose from an estimation of the 2D pose: One form of prior information adopted in this setting is the joint angle limits [Parameswaran 2004, Barrón 2001]. With the availability of mocap data, such prior information is formed in a data-driven manner [Ramakrishna 2012, Akhter 2015]. [Simo-Serra 2012, Simo-Serra 2013] presents an approach where noisy samples are predicted, which are disambiguated using kinematic constraints. [Ramakrishna 2012, Wang 2014, Zhou 2017] propose sparse bases that handle articulated deformation of human bodies that cannot be captured by PCA as well. Recent works of [Martinez 2017, Zhao 2018b] show that a mapping from 2D to 3D pose can be learned via neural networks, leading to simple baselines .

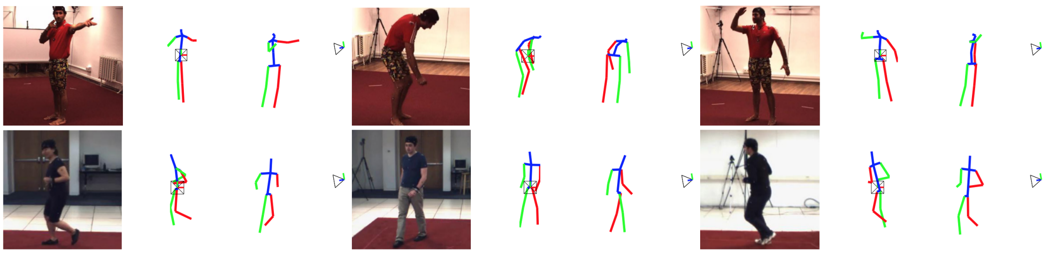


Figure 1.8: Example results for monocular 3D pose estimation. The images are from the Human 3.6M dataset [Ionescu 2014b] (top row) and HumanEva dataset [Sigal 2010] (bottom row). The results are obtained with the volumetric regression system of [Pavlakos 2017]. *Figure taken from the paper cited in the caption.*

Discriminative Learning of Monocular 3D Pose Estimation: Similar to other tasks reviewed so far, DPM based approaches such as [Sigal 2012, Belagiannis 2014] are replaced with discriminative methods, for instance using regression forests [Pons-Moll 2014, Ionescu 2014a]. Following the success of deep learning [Toshev 2014] for the 2D pose, [Li 2014, Li 2015, Tekin 2016] propose CNN-based direct regression of 3D joints. [Zhou 2016b] proposes regression of the kinematic tree. [Chen 2017a] proposes a nearest neighbor search given an estimate 2D pose from a library of projected 3D poses. [Tome 2017] proposes fusing a probabilistic model of 3D poses with a multi-stage CNN architecture and uses plausible 3D poses to improve 2D localization. [Rogez 2017] introduces the Localization-Classification-Regression system, where pose proposals are classified and further refined via regression similar to Fast-RCNN [Girshick 2015]. [Sun 2018] proposes regression of bones instead of joints, using which the 3D pose is composed. [Pavlakos 2017] proposes the volumetric regression of 3D heatmaps using CNNs, reporting improved performance. Results of this system are demonstrated in Fig. 1.8. Volumetric regression is further improved by replacing the argmax operation for localizing the center of the heatmap with soft-argmax, as shown in [Sun 2017].

Generalization to Images In-The-Wild: The mocap datasets are recorded in a studio environment. There is a domain shift problem when the trained systems are operating on real-life images with arbitrary backgrounds and occlusions. Example images from two mocap datasets can be observed in Fig. 1.8. Additionally, the number of different human bodies in training and test sets are limited, e.g. 5 different bodies in the Human 3.6m training set [Ionescu 2014b]. 2D keypoint supervision from diverse everyday life settings, e.g. the MPII dataset [Andriluka 2014] or the COCO dataset [Lin 2014], is adopted by recent works and is shown to be beneficial in terms of performance [Chen 2016a, Tekin 2017, Pavlakos 2017, Sun 2018]. [Rogez 2016] creates synthetic 3D data on real images by making analogies from mocap data based on local 2D pose similarity. There are also works that automatically synthesize semi-photorealistic images of people rendered from 3D sequences of human motion capture data [Chen 2016c, Varol 2017]. The Human3.6M dataset [Ionescu 2014b] also provides renders of people in mixed reality settings, though much limited in terms of variability and scale with respect to [Varol 2017]. The domain shift still exists with synthetic data, and it is not possible to evaluate the performance in-the-wild. The recent work of [von Marcard 2018] uses Inertial Measurement Units (IMUs) and a camera to obtain in-the-wild 3D poses. This dataset for the first time allows measuring the performance of 3D pose estimation systems in-the-wild.

1.2.2 Deformable Templates: Model-based, Top-down techniques

So far we have emphasized the power of CNN-based, bottom-up approaches in a number of computer vision problems. These tasks are essential parts of understanding the objects, but in isolation they are not descriptive. For instance, localizing some landmarks of an object alone does not allow reasoning on how the object relates to other objects of the same class. On the contrary, via top-down modelling, prior knowledge about the object's appearance, shape, part configuration can be used to better understand the characteristics of a given object instance.

Deforming templates to model different instances of the same object is an idea that has been used for centuries. Albrecht Dürer was working on deformable templates in the German Renaissance. In his work *Four Books on Human Proportion* [Durer 1534], he used fixed appearance images, which can be seen as canonical templates, and warped them with different grids to model human proportions. An example is visualized in Fig. 1.9.a, where a human face figure is transformed. Motivated from Dürer's works, D'Arcy Thompson also adopted the deformable template paradigm in his seminal work on morphogenesis, *On Growth and Form* [Thompson 1942]. In Fig. 1.9.b,c we demonstrate how he modelled different species using simple geometric transformations and a template.

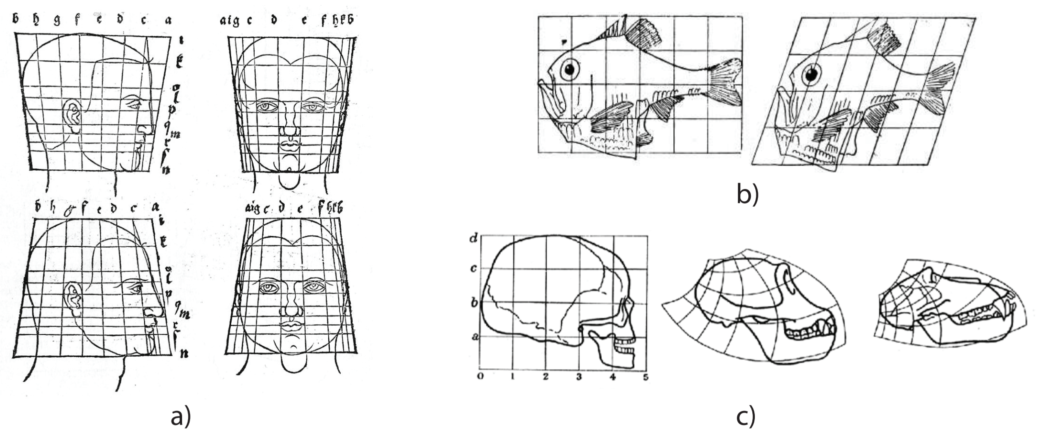


Figure 1.9: a) The use of deformable templates in the works of Albrecht Dürer [Durer 1534]. By applying simple geometric transforamtions to the template, new faces are obtained b,c) Works of D'Arcy Thompson on mathematical biology [Thompson 1942]. b) Transformation of *Argyropelecus olfersi* into *Sternoptyx diaphana* by a horizontal shear. c) Simple non-rigid geometric transformations between the skulls of a human, chimpanzee and a baboon. *Figures taken from the respective papers in the citations provided.*

Parametric models typically model the shape using characteristic deformations for a given object. [Yuille 1992] proposed detection of facial features such as the eye and mouth, using templates obtained by circles and curves in a parametric manner. [Staib 1992] used a parametric representation obtained by elliptical Fourier descriptors to represent curves. Template based deformable models [Grenander 1976] involve a prototype object that is deformed using parametric transformations to fit an observed object. For instance, [Amit 1991] used an image based hand prototype. Deformable part models [Fischler 1973, Burl 1998, Felzenszwalb 2005, Felzenszwalb 2010], introduced as useful tools for many visual tasks in the previous section

Sec. 1.2.1, are also prototype-based deformable models. Here, we follow on reviewing methods that explicitly model continuous geometric transformations.

The active shape model (ASM) [Cootes 1992] uses a collection of samples to statistically estimate an average shape for an object class. The modes of deformation of the template are modelled linearly using PCA. [Jain 1996] proposes warping templates using radial basis functions (RBF) to bring the template in alignment with the object in the image. The active appearance model (AAM) [Edwards 1998, Cootes 2001, Matthews 2004] models the shape and appearance of a deformable object class. The shape is modelled using the 'Point Distribution Model', as in the active shape models. The appearance is represented using the intensities in the template coordinate system. The appearance is modelled linearly using PCA following eigenfaces [Sirovich 1987, Turk 1991]. Another line of work that models appearance is morphable models [Vetter 1997c, Vetter 1997a, Jones 1998]. 3D morphable models (3DMMs) [Blanz 1999, Blanz 2003a] deal with the 3D shape of the object. 3D scans are utilized to learn shape bases as deviations from the mean 3D shape. Once the 3DMM is fitted, one can render the object from a different global pose or change the illumination, as shown in [Blanz 1999].

Fitting deformable models, such as AAMs is done by searching for shape and appearance parameters that maximize the matching of intensities between the model and the object in the input image. This fitting is a non-linear optimization problem. When AAMs were initially proposed [Cootes 2001], the fitting was formulated as an iterative procedure with incremental additive updates to the shape and appearance coefficients. At each iteration, the input image can be warped into the template domain to compute the error term. The cost function is similar to the one of Lucas-Kanade for affine image alignment. [Matthews 2004] proposes the inverse compositional image alignment algorithm, significantly augmenting the speed and quality of fitting. Another highly influential method to fit AAMs is 'supervised descent' [Xiong 2013], a supervised regression method. The parameters of the statistical shape model are directly regressed from image features using a cascaded architecture.

There is a large quantity of recent works that propose semantic alignment between two images [Kim 2013, Zhou 2015, Bristow 2015, Ham 2016, Zhou 2016a, Han 2017, Kim 2017b, Rocco 2017, Rocco 2018]. These methods do not find correspondences to a fixed canonical coordinate system, which would provide a more sophisticated understanding of geometry. There are works in the previous decade that aimed at learning shape/appearance factorizations in an unsupervised manner, exploring groupwise image alignment [Frey 2003, Learned-Miller 2006, Kokkinos 2007]. [Cashman 2013] proposes learning a 3D morphable model from a collection of 2D pictures annotated with few landmarks and the silhouette information. Their system works as long as the object class is not articulated and given that there is a rigid 3D model to initialize the mean shape. Recently using CNN based systems, [Thewlis 2017] uses the equivariance principle to align sets of images to a common coordinate system. Also, [Kanazawa 2018c] shows that using segmentations, landmarks and symmetry assumption one can form a 3D morphable model of an object from an image collection and demonstrates results on birds.

1.2.2.1 Deformable models of the face

Modelling the human face is critical for many computer vision applications. Also, the geometry of the face is simple with no articulations, making it straightforward to parameterize in a template space. Perhaps due to these reasons, the research on deformable templates has been driven by works focusing on modelling the face. Seminal examples would be ASMs [Cootes 1992], AAMs [Cootes 2001] and 3DMMs [Blanz 1999]. As state-of-the-art AAMs provide effective methods for alignment of faces, e.g. [Trigeorgis 2016], they do not provide a 3D understanding of the face geometry. 3DMMs, on the other hand, effectively reconstruct the face shape from in-the-wild RGB images or noisy RGBD point clouds. The first and the recent 3DMMs of the human face are demonstrated in Fig. 1.10.

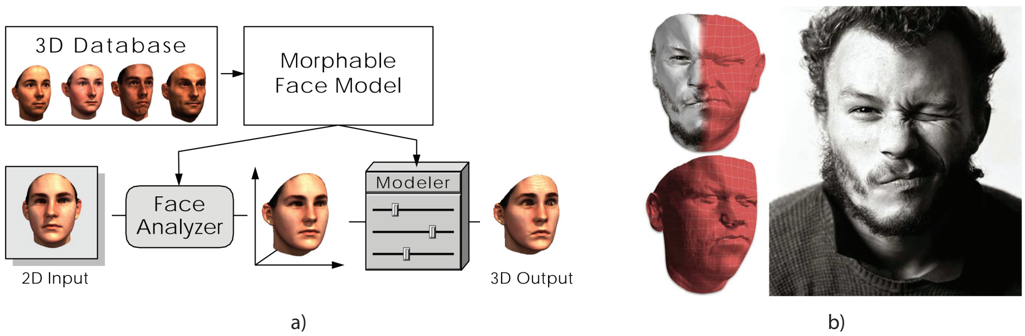


Figure 1.10: Demonstration of 3D Morphable Models (3DMMs) of the human face. a)The first 3DMM of the human face [Blanz 1999] b) The recent 3DMM of the human face learned from 10000 facial identities [Booth 2016]. *Figures taken from the respective papers in the citations provided.*

Learning 3DMMs: 3D scans of faces are used to learn 3DMMs. The most challenging step of learning the model is bringing the scanned faces in correspondence with the template. Initially, [Blanz 1999] solved the dense correspondence problem by flattening the 3D face surface. Correspondences are declared using optical flow in the flattened 2D space. [Amberg 2007] proposes learning expressions by learning a new linear subspace for deviations from the neutral pose. This allows modelling identities and expression together. [Patel 2009] proposes manual annotations of fixed face U-V coordinates, which are utilized to co-register the meshes. This supervised approach is shown to be more robust with respect to optical flow. [Paysan 2009] collects manually placed landmarks and used Non-Rigid Iterative Closest Point algorithm to align scanned faces. Their 'Basel Face Model' consists of 200 scanned subjects. [Cao 2014] captures the variability in the expression space using blendshapes. More recently, [Booth 2016] proposes a 3DMM automatically constructed from scans of 10000 different facial identities, covering diverse age and ethnicity groups.

Fitting 3DMMs: The initial 3DMM fitting approach was via analysis-by-synthesis-based optimization. The proposed fitting approach [Blanz 1999] was minimizing appearance differences via stochastic gradient descent. The following work of [Romdhani 2005] utilized more sophisticated features to define the objective function. Recently, [Schönborn 2017] reports improved results via probabilistic interpretation using Markov Chain Monte Carlo.

Recent works on fitting 3DMMs mostly rely on the power of CNNs within discriminative frameworks. [Zhu 2016] proposes an iterative approach to estimate the model parameters. The input to their iterative CNN system is the image and a rendered representation from the previous iteration. [Huber 2016] describes a cascaded method that is based on landmark regression. [Jourabloo 2016] uses landmarks to fit a 3DMM. They train a CNN to regress pose and shape parameters of the fitted 3DMMs. [Richardson 2016] exploits synthetic data to train an iterative network. [Tran 2017] proposes a system where the same shape parameters are enforced for different images of the same subject. [Kim 2017a] incorporates illumination parameters by inverse rendering and train on synthetic images. [Jackson 2017] proposes to regress the shape in a voxelized volume. [Sela 2017] uses an FCN to predict correspondences and depth, which are used to improve the quality of the fit. [Bas 2017] proposes the use of the 3D morphable model as a spatial transformer network that outputs a flattened 2D texture space.

1.2.2.2 Deformable models of the human body

We have observed in the previous section, Sec. 1.2.1, how deformable part models were popular in human understanding tasks, such as detection [Felzenszwalb 2008], pose estimation [Felzenszwalb 2005] and 3d pose estimation [Belagiannis 2014]. There is a rich literature regarding top-down 3D understanding of human motions from videos. In their seminal work, [Marr 1978], proposed a compositional 3D shape representation for the human body. [Hogg 1983] worked on model-based analysis-by-synthesis methodology. Many following works have used part based 3D models for recognition of 3D human motions from videos [Rohr 1994, Gavrila 1996, Ju 1996, Sidenbladh 2000, Duetscher 2000, Kakadiaris 2000, Sminchisescu 2003, Sigal 2004]. Such manually designed models are now replaced with those learned from scan data. Evolution of human body models is depicted in Fig. 1.11 with several examples from different decades: the cylinder based hierarchical model of [Marr 1978], ellipsoid based [Gavrila 1996], models learned from scans of actual humans such as SCAPE and SMPL [Anguelov 2005, Loper 2015].

There are some challenges involved in obtaining morphable models of the human body based on 3D scanned examples similar to the morphable face model [Blanz 1999]. Due to articulations, it is not straightforward to align the 3D human body shapes to model shape variations. The common approach is to bring

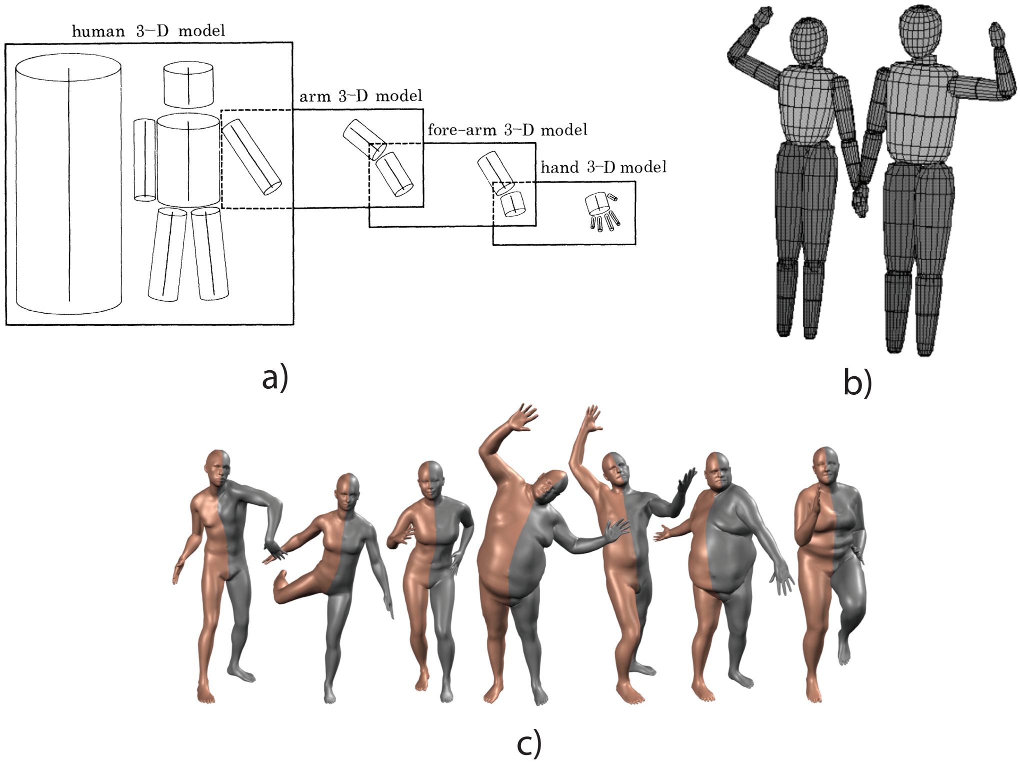


Figure 1.11: Evolution of human body models with several examples. a) The cylinder based hierarchical model of [Marr 1978]. b) [Gavrila 1996] model with ellipsoid parts. c) [Loper 2015] model based on human body scans. *Figures taken from the respective papers in the citations provided.*

the scans into the same pose by modelling or learning how vertices are associated with a hand-engineered skeleton structure. This is known as 'skeleton subspace deformation modelling' or 'blend skinning'. In linear blend skinning, vertices are transformed using a weighted influence of the bones associated with them. There are common artefacts near the joints such as stretching and undesired protrusions. To cope with these issues, [Lewis 2000] proposes 'pose space deformation (PSD) model', where extra deformations are defined as a function of the joint angles. Manual modelling of these deformations can be considered as the current standard practice for gaming and animations. Starting with the work of [Allen 2002], numerous works learn PSD models from scan data, e.g. work of [Kry 2002] on modelling scanned hands.

The first methods to characterize the space of human body shapes linearly was [Allen 2003] using the CEASAR dataset [Robinette 1999]. [Seo 2003] analyzes the deformation using rigid with non-rigid components modelled using PCA. This was followed by the SCAPE model [Anguelov 2005], who models the pose deformation as a function of the pose of the articulated skeleton for the first time in this context. Following SCAPE, several other methods used triangle deformations [Hasler 2009, Hirshberg 2012, Chen 2013]. [Hasler 2010] proposes using bones to model the shape of the human body. [Pons-Moll 2015a] proposed Dyna, where a 4D capture system is used to scan soft-tissue deformations in time and a low-dimensional linear subspace approximating this soft-tissue deformation is learned. [Allen 2006] proposes modelling identity-dependent and pose-dependent shape variation in a correlated fashion. The vertices are modelled in the rest pose, referred to as dress shape. Corrective deformations, dealing with skinning artifacts are also applied in this space. Following [Allen 2006], [Loper 2015] proposes a simpler model, SMPL, where pose blendshapes are regressed from a vector of concatenated part relative rotation matrices defined on joint angles. The authors of [Loper 2015] argue that their simpler modelling makes training easier and their model generalizes better since more samples are used. An advantage of the SMPL model with respect to existing deformable models of the human body is its compatibility with graphics tools and game engines. The stitched puppet [Zuffi 2015] model, represents the human body by a graphical model of parts that can translate and rotate in 3D independently. The model deforms to represent different body shapes and to capture pose-dependent shape variations. Recently, [Hesse 2018] proposes SMIL, 3D Skinned Multi-Infant Linear body model from noisy and incomplete RGB-D data, which could be instrumental in the detection of developmental disorders. [Joo 2018] proposes markerless capture of facial expressions, body motion, and hand gestures. They propose learning a detailed deformable model by locally stitching together models of hand and face to the body. They also learn a new unified model, called Adam, by sampling instances of the stitched model.

Fitting 3D Human Models: There are recent efforts in fitting the 3DMM of the human body to monocular images. The SMPL model [Loper 2015] is adopted in these methods. In 'SMPLify' [Bogo 2016], the pose and shape parameters of the SMPL model are optimized along with camera parameters such that the keypoints on the model are in alignment with 2d keypoints estimated using state-of-the-art keypoint estimators. This is a hard optimization problem, which often fails, as shown by [Lassner 2017b]. [Lassner 2017b] fits the model using the SMPLify method to cropped humans from natural images and asks human annotators to filter the renders. More than half of the fits are filtered out, leaving better fits to train discriminative models for segmentation and 91 landmark localization. [Pavlakos 2018, Omran 2018, Kanazawa 2018a, Zanfir 2018] propose regression of pose and shape parameters of the SMPL model using neural networks directly from the image. [Varol 2018] proposes a multi-task system that outputs the 3D shape in a voxelized space. SMPL parameters are then optimized to overlap with the estimated 3D shape. One significant limitation of these fitting approaches is the expressiveness of the existing deformable models. Current state-of-the-art models are trained using limited diversity in terms of ethnicity and age, for instance, children cannot be reconstructed using these models.

1.3 Structure of the Thesis

So far we have introduced the theme of the thesis, listed our main contributions and reviewed the relevant literature. For the rest of the thesis, the organization of the chapters follows the chronological progression of the contributions. The outline is as follows:

We firstly introduce 'dense shape regression', DenseReg, to establish dense correspondences between image pixels and a 3D template of the human face in Chapter 2. We define correspondences using a continuous, canonical parameterization of the template as in statistical deformable models in Sec. 2.2. We introduce 'quantized regression', a method that first selects a rough quantized position and then refines the localization through regression of the residuals in Sec. 2.3. We present results for facial landmark localization on images and videos, facial part segmentation and ear shape reconstruction in Sec. 2.4.

In Chapter 3 we present quantized regression and structured prediction for deep monocular 3D human pose estimation. We show that the quantized regression effectively predicts locations of human keypoints on a volumetric label space. We also adopt a structured model that imposes constraints between the relative positions of parts in Sec. 3.2. We experiment with various graphical model connectivities and report results in Sec. 3.3.

Chapter 4 introduces the task of 'dense human pose estimation', DensePose. We propose a 2D parameterization of the human body surface by flattening semantically meaningful parts. We propose an efficient system for collecting image-to-surface annotations and collect millions of manually annotated correspondences on the human body. Our annotation system and the collected dataset are presented in Sec. 4.2. We then describe our system that predicts per-instance correspondences in Sec. 4.3. We report quantitative results based on the collected annotations along with qualitative results on scenes with multiple people and occlusions in Sec. 4.4.

Chapter 5 introduces DensePose guided human pose transfer between two images. We synthesize a new image based on appearance and pose obtained from different images. The proposed two-stream system is presented in Sec. 5.2. The results are presented in Sec. 5.3, where we show that conditioning on the proposed dense human pose leads to better synthesis with respect to alternative pose representations such as sparse landmarks and body parts.

Finally, in Chapter 6, we provide concluding remarks and discuss future directions of research.

1.4 List of Publications

1. RA Guler, N Neverova, I Kokkinos. DensePose: Dense human pose estimation in-the-wild. (Oral) **CVPR 2018**
2. N Neverova, RA Guler, I Kokkinos. Dense pose transfer. **ECCV 2018**
3. Z Shu, M Sahasrabudhe, RA Guler, D Samaras, N Paragios, I Kokkinos Deforming Autoencoders: Unsupervised Disentangling of Shape and Appearance. **ECCV 2018**
4. RA Guler, G Trigeorgis, E Antonakos, P Snape, S Zafeiriou, I Kokkinos. DenseReg: Fully convolutional dense shape regression in-the-wild. **CVPR 2017**
5. S Kinauer*, RA Guler*, S Chandra, I Kokkinos. Structured Output Prediction and Learning for Deep Monocular 3D Human Pose Estimation. **EMM-CVPR 2017**
6. RA Guler, I Kokkinos et.al. Human Joint Angle Estimation and Gesture Recognition for Assistive Robotic Vision. (Oral) **ECCV Workshop 2016**

Dissemination Activities

- Supplementary materials, videos and links to our open sourced codes are presented in `https://alpguler.com`.
- DenseReg and DensePose have been presented as real-time demonstrations in CVPR 2017 and CVPR 2018 with 'texture mapping' applications.
- Two challenges co-organized in ECCV 2018, introducing the 'dense human pose esimtation' task within the COCO challenge involving static images and PoseTrack challenge involving videos.

Chapter 2

Fully Convolutional Dense Shape Regression

In this chapter we propose a system to establish dense correspondences between a 3D object model and an image "in the wild". We introduce 'DenseReg', a fully-convolutional neural network (F-CNN) that *dens*ely *reg*resses, at every foreground pixel, a pair of U-V template coordinates in a single feedforward pass.

To train DenseReg we construct a supervision signal by combining 3D deformable model fitting and 2D landmark annotations. We define the regression task in terms of the intrinsic, U-V coordinates of a 3D deformable model that is brought into correspondence with image instances at training time. A variety of other object-related tasks (e.g. part segmentation, landmark localization) are shown to be by-products of this task and to largely improve thanks to its introduction.

We obtain highly-accurate regression results by combining ideas from semantic segmentation with regression networks, yielding a 'quantized regression' architecture that first obtains a quantized estimate of position through classification and then refines it through regression of the residual.

This work was published at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017).

2.1 Introduction

We introduce a discriminatively trained network to obtain, in a fully-convolutional manner, dense correspondences between an input image and a deformation-free template coordinate system. We exploit the availability of manual landmark annotations "in-the-wild" in order to fit a 3D template; this provides us with a dense correspondence field, from the image domain to the 2-dimensional, $U-V$ parameterization of the surface. We then train a fully convolutional network that densely regresses from the image pixels to this $U-V$ coordinate space. This combines the fine-grained discrimative power of statistical deformable models with the "in the wild" operation of fully-convolutional neural networks.

We show experimentally that the proposed feedforward system outperforms substantially more involved systems developed in particular for facial landmark localization while also outperforming the results of systems trained on lower-granularity tasks, such as facial part segmentation. We can also seamlessly integrate this method with iterative, deformable model-based algorithms to obtain results that

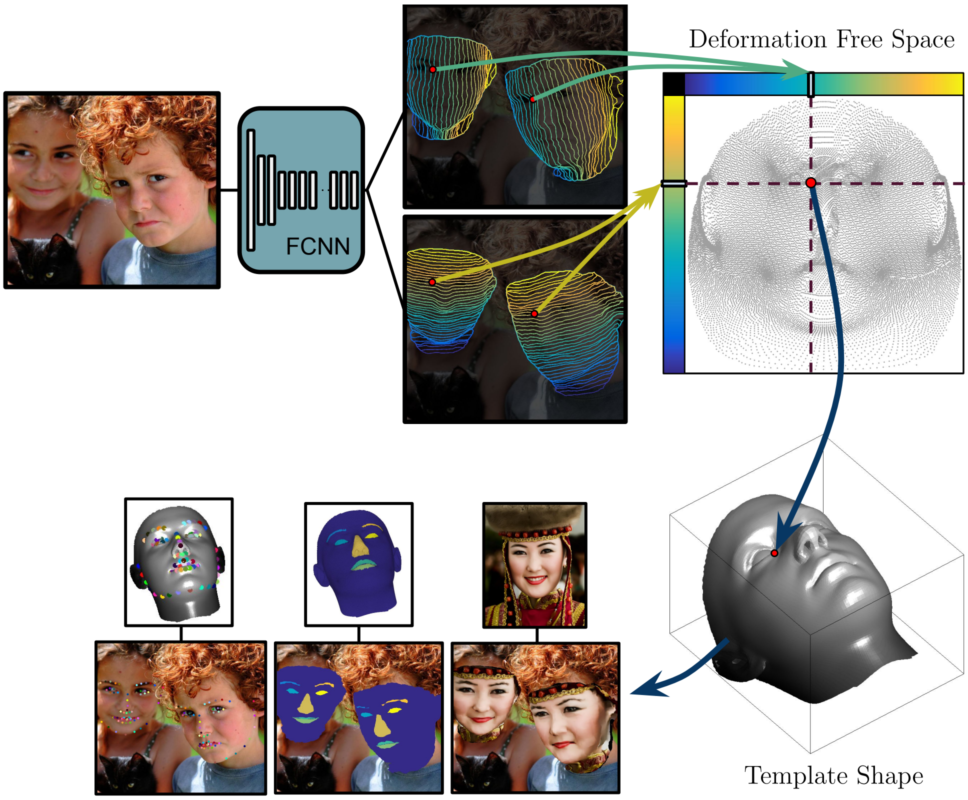


Figure 2.1: We introduce a fully convolutional neural network that regresses from the image to a “canonical”, deformation-free parameterization of the face surface, effectively yielding a dense 2D-to-3D surface correspondence field. Once this correspondence field is available, one can effortlessly solve many image-level problems by backward-warping their canonical solution from the template coordinates to the image domain for the problems of landmark localization, semantic part segmentation, and face transfer.

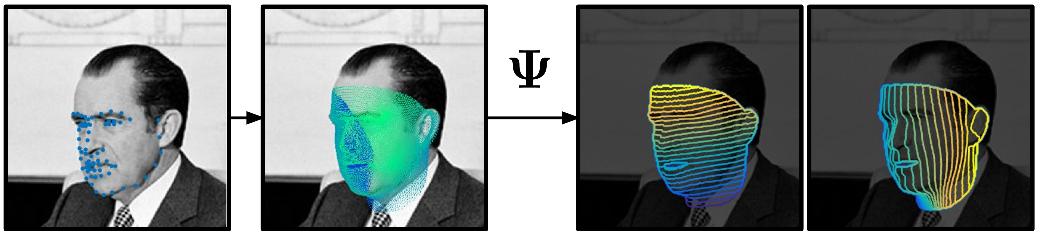
constitute the current state-of-the-art on large-scale, challenging facial landmark localization benchmarks.

We can summarize our contributions as follows:

- We introduce the task of dense shape regression in the setting of CNNs, and exploit the notion of a deformation-free UV-space to construct target ground-truth signals (Sec.2.2).

- We propose a carefully-designed fully-convolutional shape regression system that exploits ideas from semantic segmentation and dense regression networks. Our *quantized regression* architecture (Sec.2.3) is shown to substantially outperform simpler baselines that consider the task as a plain regression problem.

- We use dense shape regression to jointly tackle a multitude of problems, such as landmark localization or semantic segmentation. In particular, the tem-



(a) (b) (c)

Figure 2.2: Ground truth generation: *(a)* Annotated landmarks. *(b)* Template shape morphed based on the landmarks. *(c)* Deformation-free coordinates (u^h and u^v), obtained by unwrapping the template shape, transferred to image domain.

plate coordinates allow us to transfer to an image multiple annotations constructed on a single template system, and thereby tackle multiple problems through a single network.

- We use the regressed shape coordinates for the initialization of statistical deformable models; systematic evaluations on facial analysis benchmarks show that this yields substantial performance improvements on tasks.

- We demonstrate the generic nature of the method by applying it to the task of estimating dense correspondence in other object, such as the human ear.

2.2 From SDMs to Dense Shape Regression

Following the deformable template paradigm [Yuille 1991, Amit 1991], we consider that object instances are obtained by deforming a prototypical object, or 'template', through dense deformation fields. This makes it possible to factor object variability within a category into variations that are associated to deformations, generally linked to the object's 2D/3D shape, and variations that are associated to appearance (or, 'texture' in graphics), e.g. due to facial hair, skin color, or illumination.

This factorization largely simplifies the modelling task. SDMs use it as a stepping stone for the construction of parametric models of deformation and appearance. For instance, in AAMs a combination of Procrustes Analysis, Thin-Plate Spline warping and PCA is the standard pipeline for learning a low-dimensional linear subspace that captures category-specific shape variability [Cootes 2001]. Even though we have a common starting point, rather than trying to construct a linear generative model of deformations, we treat the image-to-template correspondence as a vector field that our network tries to regress.

In particular, we start from a template $\mathbf{X} = [\mathbf{x}_1^\top, \mathbf{x}_2^\top, ...\mathbf{x}_m^\top]^\top \in \mathbb{R}$, where each $\mathbf{x}_j \in \mathbb{R}^3$ is a vertex location of the mesh in 3D space.

This template could be any 3D facial mesh, but in practice it is most useful to use a topology that is in correspondence with a 3D statistical shape model such

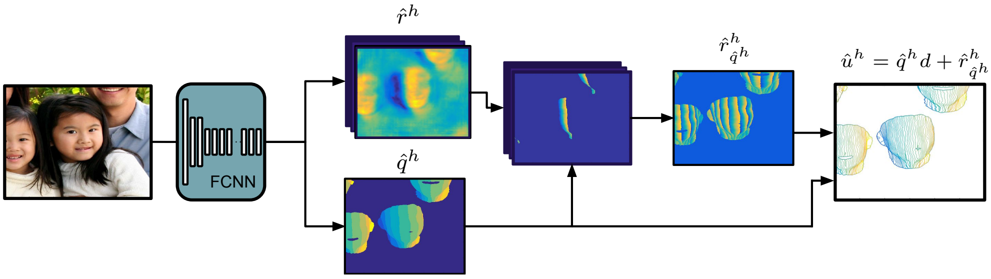


Figure 2.3: Proposed Quantized Regression Approach for the horizontal correspondence signal: The continuous signal is regressed by first estimating a grossly quantized (or, discretized) function through a classification branch. For each quantized value $\hat{q}^h$ we use a separate residual regression unit's prediction, $\hat{r}^h_{\hat{q}^h}$, effectively multiplexing the different residual predictions. These are added to the quantized prediction, yielding a smooth and accurate correspondence field.

as [Booth 2016] or [Paysan 2009]. We compute a bijective mapping ψ, from template mesh $\mathbf{X}$ to the 2D canonical space $\mathbf{U} \in \mathbb{R}^{2\times m}$, such that

$$\psi(\mathbf{x}_j) \mapsto \mathbf{u}_j \in \mathbf{U} \quad , \quad \psi^{-1}(\mathbf{u}_j) \mapsto \mathbf{x}_j. \tag{2.1}$$

The mapping ψ is obtained via the cylindrical unwrapping described in [Booth 2014]. Thanks to the cylindrical unwrapping, we can interpret these coordinates as being the horizontal and vertical coordinates while moving on the face surface: $u^h_j \in [0,1]$ and $u^v_j \in [0,1]$. Note that this semantically meaningful parameterization has no effect on the operation of our method.

We exploit the availability of landmark annotations "in the wild", to fit the template face to the image by obtaining a coordinate transformation for each vertex $\mathbf{x}_j$. We use the fittings provided by [Zhu 2016] which were fit using a modified 3DMM implementation [Romdhani 2005]. However, for the purpose of this paper, we require a per-pixel estimate of the location in UV space on our template mesh and thus do not require an estimate of the projection or model parameters as required by other 3D landmark recovery methods [Jourabloo 2016, Zhu 2016]. The per-pixel UV coordinates are obtained through rasterization of the fitted mesh and non-visible vertices are culled via z-buffering.

As illustrated in Fig. 2.2, once the transformation from the template face vertices to the morphed vertices is established, the $\mathbf{u}_j$ coordinates of each visible vertex on the canonical face can be transferred to the image space. This establishes the ground truth signal for our subsequent regression task.

2.3 Fully Convolutional Dense Shape Regression

Having described how we establish our supervision signal, we now turn to the task of estimating it through a convolutional neural network (CNN). Our aim is to estimate at any image pixel that belongs to a face region the values of $\mathbf{u} = [u^h, u^v]$. We need to also identify non-face pixels, e.g. by predicting a 'dummy' output.

One can phrase this problem as a generic regression task and attack it with the powerful machinery of CNNs. Unfortunately, the best performance that we could obtain this way was quite underwhelming, apparently due to the task's complexity. Our approach is to quantize and estimate the quantization error separately for each quantized value. Instead of directly regressing u, the quantized regression approach lets us solve a set of easier sub-problems, yielding improved regression results.

In particular, instead of using a CNN as a 'black box' regressor, we draw inspiration from the success of recent works on semantic part segmentation [Tsogkas 2015, Chen 2018b], and landmark classification [Newell 2016]. These works have shown that CNNs can deliver remarkably accurate predictions when trained to predict *categorical variables*, indicating for instance the facial part or landmark corresponding to each pixel.

Building on these successes, we propose a hybrid method that combines a classification with a regression problem. Intuitively, we first identify a coarser face region that can contain each pixel, and then obtain a refined, region-specific prediction of the pixel's $U - V$ field. As we will describe below, this yields substantial gains in performance when compared to the baseline of a generic regression system.

For the human bodies, the regions are modeled by hand and for the facial regions, we use a simple geometric approach: We tessellate the template's surface with a cartesian grid, by uniformly and separately quantizing the u^h and u^v coordinates into K bins, where K is a design parameter. For any image that is brought into correspondence with the template domain, this induces a discrete labelling, which can be recovered by training a CNN for classification.

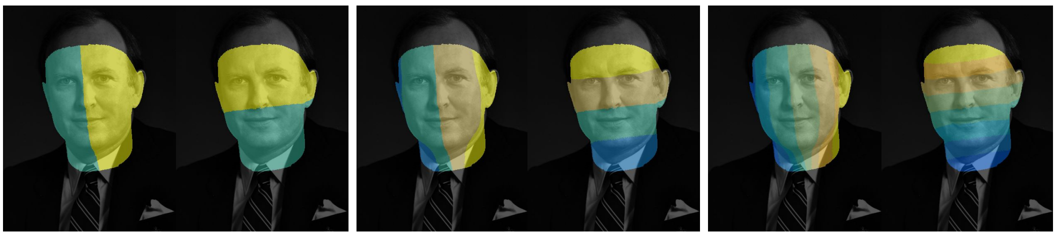


Figure 2.4: Horizontal and vertical tessellations obtained using $K = 2, 4$ and 8 bins.

On Fig. 2.4, the tesselations of different granularities are visualized. For a sufficiently large value of K even a plain classification result could provide a reasonable estimate of the pixel's correspondence field, albeit with some staircasing effects. The challenge here is that as the granularity of these discrete labels becomes increasingly large, the amount of available training data decreases and label complexity increases.

We propose to combine powerful classification results with a regression problem that will yield a refined correspondence estimate. For this, we compute the residual between the desired and quantized $U - V$ coordinates and add a separate module that tries to regress it. We train a separate regressor per facial region, and at any pixel only penalize the regressor loss for the responsible face region. We can interpret this form as a 'hard' version of a mixture of regression experts [Jordan 1994].

The horizontal and vertical components u^h, u^v of the correspondence field are predicted separately. This results in a substantial reduction in computational and sample complexity - For K distinct U and V bins we have K^2 regions; the classification is obtained by combining 2 K-way classifiers. Similarly, the regression mapping involves K^2 regions, but only uses $2K$ one-dimensional regression units. The pipeline for quantized face shape regression is provided in Fig. 2.3.

We now detail the training and testing of this network; for simplicity we only describe the horizontal component of the mapping. From the ground truth construction, every position $\mathbf{x}$ is associated with a scalar ground-truth value u^h. Rather than trying to predict u^h as is, we transform it into a pair of discrete q^h and continuous r^h values, encoding the quantization and residual respectively:

$$q^h = \lfloor \frac{u^h}{d} \rfloor, \quad r_i^h = \left(u_i^h - q_i^h d \right), \tag{2.2}$$

where $d = \frac{1}{K}$ is the quantization step size (we consider u^h, u^v coordinates to lie in $[0, 1]$).

Given a common CNN trunk, we use two classification branches to predict q^h, q^v and two regression branches to predict r^h, r^v as convolution layers with kernel size 1×1. As mentioned earlier, we employ separate regression functions per region, which means that at any position we have K estimates of the horizontal residual vector, $\hat{r}_i^h, \ i = 1, \ldots, K$.

At test time, we let the network predict the discrete bin $\hat{q}^h$ associated with every input position, and then use the respective regressor output $\hat{r}^h_{\hat{q}^h}$ to obtain an estimate of u:

$$\hat{u}^h = \hat{q}^h d + \hat{r}^h_{\hat{q}^h} \tag{2.3}$$

For the q^h and q^v, which are modeled as categorical distributions, we use softmax followed by the cross entropy loss. For estimating $\hat{r}^h$ and $\hat{r}^v$, we use a normalized version of the smooth L_1 loss [Girshick 2015]. The normalization is obtained by dividing the loss by the number of pixels that contribute to the loss.

2.3.1 Quantized Regression as Mixture of Experts

In our formulation, $\hat{q}^h$ is modeled using a categorical distribution and is trained using softmax followed by cross entropy loss. This reconstruction can also be seen as:

$$\hat{u}^h = \sum_{i=0}^{K-1} 1_{(\hat{q}^h = i)} (i \cdot d + \hat{r}_i^h), \tag{2.4}$$

where $(i \cdot d + \hat{r}_i^h)$ is the reconstruction by the i_{th} regressor and $1_{(\hat{q}^h = i)}$ is an indicator function, determining when the i_{th} regressor is active. Note that $i \cdot d$ is the value of $\hat{q}^h$, where i_{th} regressor is active.

Instead of this hard quantization, one can use a soft-quantization using the softmax function as:

$$\hat{u}^h = \sum_{i=0}^{K-1} \left(\frac{e^{f_i^{q^h}}}{\sum_j e^{f_j^{q^h}}} \right) (i \cdot d + \hat{r}_i^h), \tag{2.5}$$

where f^{q^h} is the output of the CNN branch trained for the quantized ($\hat{q}^h$) field. Notice that this is the *mixture of experts* model, [Jordan 1994], where the soft-quantization is analogous to the output of the gating network. It is straightforward to change our model accordingly: shifting each $\hat{r}_i^h$ by adding $(i \cdot d)$ to the bias terms of the corresponding 1×1 convolutional layer and weighting each 'locally trained regressor' output by the softmax function and summing up.

2.3.2 Effect of Quantization to Regression Performance

Compared to plain regression of the coordinates, the proposed quantized regression method achieves significantly better results. In Fig. 2.5 we report results of an experiment that evaluates the contribution of the q-r branches separately for different granularities. The results for the quantized branch are evaluated by transforming the discrete horzintal/vertical label into the center of the region corresponding to the quantized horizontal/vertical value respectively. The results show the merit of adopting the classification branch, as the finely quantized results(K=40,60) yield better coordinate estimates with respect to the non-quantized alternative (K=1). After K=40, we observe an increase in the failure rate for the quantized branch. The experiment reveals that the proposed quantized regression outperforms both *non-quantized* and the best of *only-quantized* alternatives. For the human shape, the partitioning can be considered as the quantization.

2.4 Experiments

Herein, we evaluate the performance of the proposed method (referred to as `DenseReg`) on various face-related tasks.

In the following sections, we first describe the training setup (Sec. 2.4.1) and then present extensive quantitative results on *(i)* semantic segmentation (Sec. 2.4.2), *(ii)* landmark localization on static images (Sec. 2.4.3), *(iii)* deformable tracking (Sec. 2.4.4), *(iv)* monocular depth estimation (Sec. 2.4.5) and *(vi)* human ear landmark localization (Sec. 2.4.5.1).

Due to space constraints, we refer to the supplementary material for additional qualitative results, experiments on monocular depth estimation and further analysis of experimental results.

2.4.1 Training Setup

Training Databases

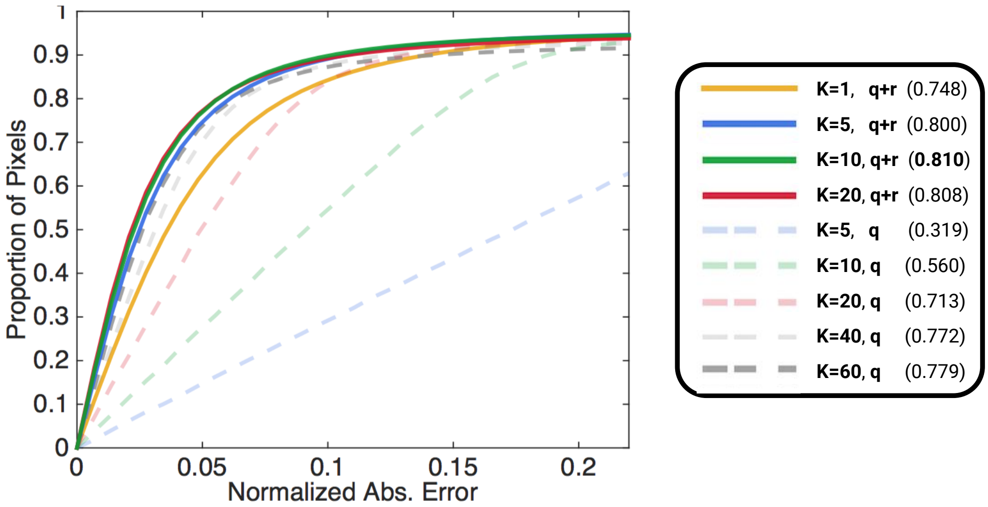


Figure 2.5: Performance of q and r, branches for various tesselation granularities of the human face, K. Areas under the curve(AUC) are reported.

We train our system using the 3DDFA data of [Zhu 2016]. The 3DDFA data provides projection and 3DMM model parameters for the Basel [Paysan 2009] + FaceWarehouse [Cao 2014] model for each image of the 300W database. We use the topology defined by this model to define our UV space and rasterize the images to obtain per-pixel ground truth UV coordinates. Our training set consists of the LFPW trainset, Helen trainset and AFW, thus 3148 images that are captured under completely unconstrained conditions and exhibit large variations in pose, expression, illumination, age, etc. Many of these images contain multiple faces, some of which are not annotated. We deal with this issue by employing the out-of-the-box DPM face detector of Mathias et al. [Mathias 2014] to obtain the regions that contain a face for all of the images. The detected regions that do not overlap with the ground truth landmarks do not contribute to the loss. For training and testing, we have rescaled the images such that their largest side is 800 pixels.

CNN Training for DenseReg

We have used two different network architectures for our experiments. In particular, in order to be directly comparable to the DeepLab-v2 network in semantic segmentation experiments we first used a ResNet101 [He 2016] architecture with dilated convolutions (atrous) [Chen 2018b], such that the stride of the CNN is 8 and (b) an Hourglass-type network [Newell 2016]. We use bilinear interpolation to upscale both the $\hat{q}$ and $\hat{r}$ branches before the losses. The losses are applied at the input image scale and back-propagated through interpolation. We apply a weight to the smooth $L1$ loss layers to balance their contribution. In our experiments, we have used a weight of 40 for quantized ($d = 0.1$) and a weight of 70 for non-quantized regression, which are determined by a coarse cross validation.

For the dense regression network, we adopt a ResNet101 [He 2016] architecture with dilated convolutions (atrous) [Chen 2018b], such that the stride of the CNN is 8. We use bilinear interpolation to upscale both the $\hat{q}$ and $\hat{r}$ branches before the losses. The losses are applied at the input image scale and back-propagated through interpolation. We apply a weight to the smooth $L1$ loss layers to balance their contribution. In our experiments, we have used a weight of 40 for quantized ($d = 0.1$) and a weight of 70 for non-quantized regression, which are determined by a coarse cross validation. We initialize the training with a network pre-trained for the MS COCO segmentation task [Lin 2014]. The new layers are initialized with random weights drawn from Gaussian distributions. Large weights of the regression losses can be problematic at initialization even with moderate learning rates. To cope with this, we use initial training with a lower learning rate for a *warm start* for a few iterations. We then use a base learning rate of 0.001 with a polynomial decay policy for $20k$ iterations with a batch size of 10 images. During training, each sample is randomly scaled with one of the ratios $[0.5, 0.75, 1, 1.25, 1.5]$ and cropped to form a fixed 321×321 input image.

2.4.2 Semantic Segmentation

As discussed in Sec. 2.2, any labelling function defined on the template shape can be transferred to the image domain using the regressed coordinates. One application that can be naturally represented on the template shape is semantic segmentation of facial parts. To this end, we manually defined a segmentation mask of 8 classes (right/left eye, right/left eyebrow, upper/lower lip, nose, other) on the template shape, as shown in Fig. 2.6.

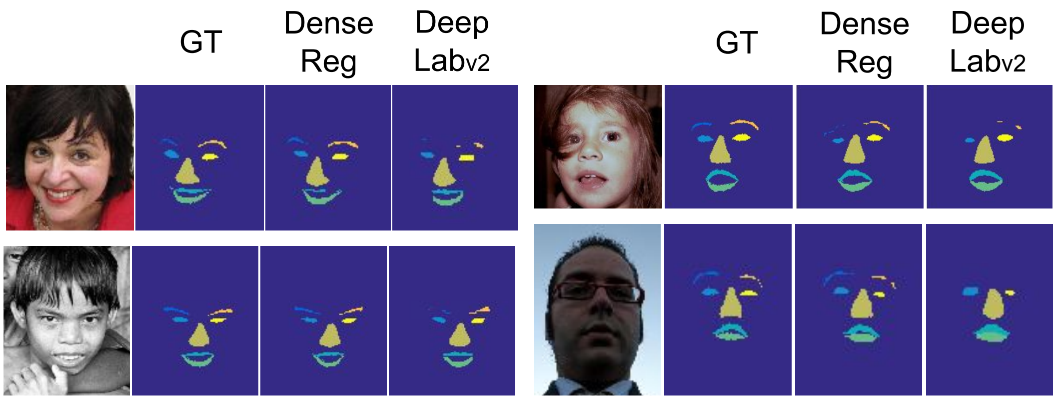


Figure 2.6: Example semantic segmentation results.

We compare against a state-of-the-art semantic part segmentation system (DeepLab-v2) [Chen 2018b] which is based on the same ResNet-101 architecture as our proposed DenseReg. We train DeepLab-v2 on the same training images (i.e. LFPW trainset, Helen trainset and AFW). We generate the ground-truth segmentation labels for both training and testing images by transferring the segmentation mask using the ground-truth deformation-free coordinates explained in Sec. 2.2. We employ the Helen testset [Le 2012] for the evaluation.

Table 2.1 reports evaluation results using the intersection-over-union (IoU) ratio. Additionally, Fig. 2.6 shows some qualitative results for both methods, along with the ground-truth segmentation labels. The results indicate that the DenseReg outperforms DeepLab-v2. The reported improvement is substantial for several parts, such as eyebrows and lips. We believe that this result is significant given that DenseReg is not optimized for the specific task-at-hand, as opposed to DeepLab-v2 which was trained for semantic segmentation. This performance difference can be justified by the fact that DenseReg was exposed to a richer label structure during training, which reflects the underlying variability and structure of the problem.

Class	*Methods*	
	DenseReg	Deeplab-v2
Left Eyebrow	48.35	40.57
Right Eyebrow	46.89	41.85
Left Eye	75.06	73.65
Right Eye	73.53	73.67
Upper Lip	69.52	62.04
Lower Lip	75.18	70.71
Nose	87.71	86.76
Other	99.44	99.37
Average	**71.96**	68.58

Table 2.1: Semantic segmentation accuracy on Helen testset measured using intersection-over-union (IoU) ratio.

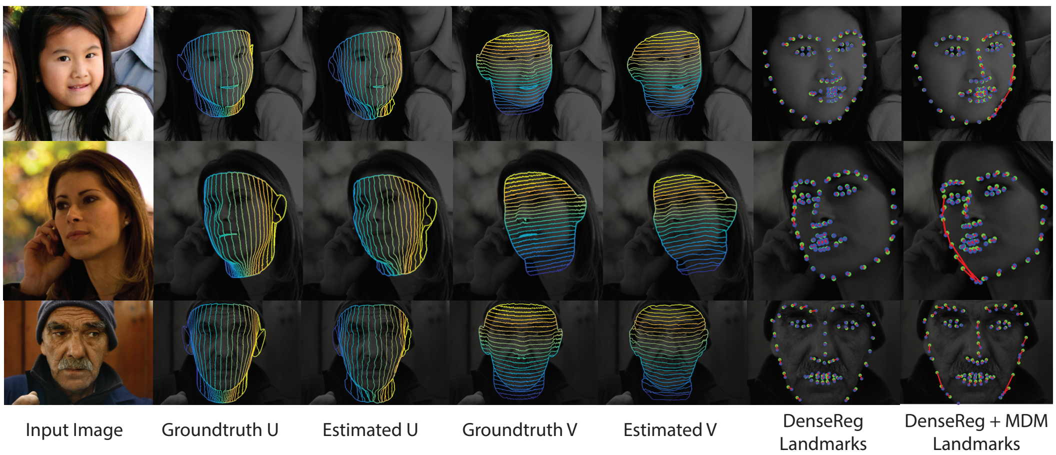


Figure 2.7: Qualitative Results. Ground-truth and estimated deformation-free coordinates and landmarks obtained from DenseReg and DenseReg+MDM are presented. Estimated landmarks(blue), ground-truth(green), lines between estimated and ground-truth landmarks(red).

2.4.3 Landmark Localization on Static Images

DenseReg can be readily used for the task of facial landmark localization on static images. Given the landmarks' locations on the template shape, it is straightforward to estimate the closest points in the deformation-free coordinates on the images. The local minima of the Euclidean distance between the estimated coordinates and the landmark coordinates are considered as detected landmarks. In order to find the local minima, we simply analyze the connected components separately. Even though more sophisticated methods for covering "touching shapes" can be used, we found that this simplistic approach is sufficient for the task.

Note that the closest deformation-free coordinates among all *visible* pixels to a landmark point is not necessarily the correct corresponding landmark. This phenomenon is called "landmark marching" [Zhu 2015] and mostly affects the jaw landmarks which are dependent on changes in head pose. It should be noted that we do not use any explicit supervision for landmark detection nor focus on ad-hoc methods to cope with this issue. Errors on jaw landmarks due to invisible coordinates

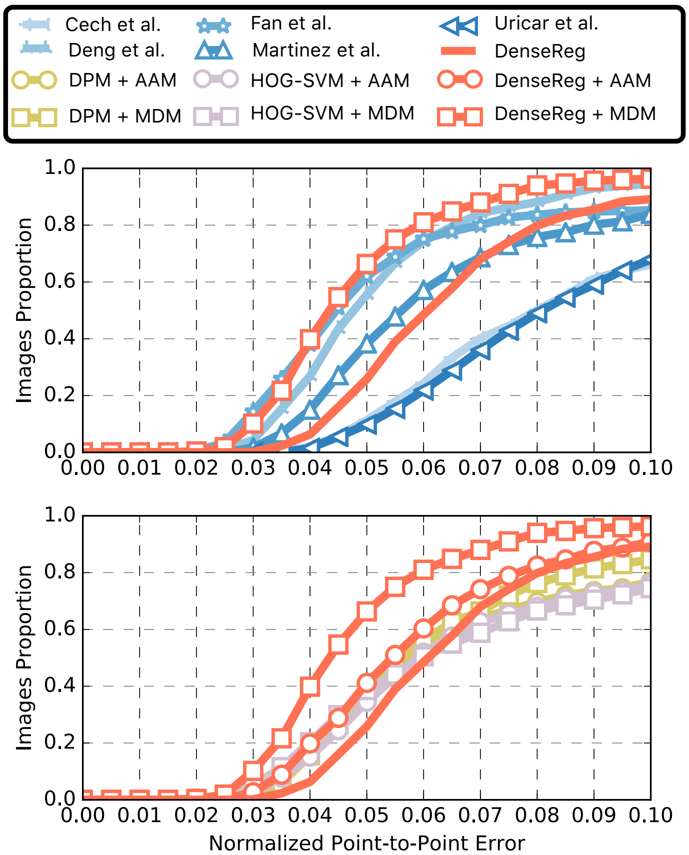


Figure 2.8: Landmark localization results on the 300W testing dataset using 68 points. Accuracy is reported as Cumulative Error Distribution of RMS point-to-point error normalized with interocular distance. *Top:* Comparison with state-of-the-art. *Bottom:* Self-evaluation results.

and improvements thanks to deformable models can be observed in Fig. 2.7.

Herein, we evaluate the landmark localization performance of DenseReg as well as the performance obtained by employing DenseReg as an initialization for deformable models [Papandreou 2008, Tzimiropoulos 2014, Antonakos 2015, Trigeorgis 2016] trained for the specific task. We present experimental results using the challenging 300W benchmark. This is the testing database that was used in the 300W competition [Sagonas 2013, Sagonas 2016] - the most important facial landmark localization challenge. The error is measured using the point-to-point RMS error normalized with the interocular distance and reported in the form of Cumulative Error Distribution (CED). Figure 2.8 (bottom) presents some self-evaluations in which we compare the quality of initialization for deformable modelling between DenseReg and two other standard face detection techniques (HOG-SVM [King 2015], DPM [Mathias 2014]). The employed deformable models are the popular generative approach of patch-based Active Appearance Models (AAM) [Papandreou 2008, Tzimiropoulos 2014, Antonakos 2015], as well as the current state-of-the-art approach of Mnemonic Descent Method (MDM) [Trigeorgis 2016]. It is interesting to notice that the performance of DenseReg without any additional deformable model on top, already outperforms even HOG-SVM detection combined with MDM. Especially when DenseReg is combined with MDM, it greatly outperforms all other combinations.

Method	*AUC*	*Failure Rate (%)*
DenseReg + MDM	**0.5219**	**3.67**
DenseReg	0.3605	10.83
Fan et al. [Fan 2016]	0.4802	14.83
Deng et al. [Deng 2016]	0.4752	5.5
Martinez et al. [Martinez 2016]	0.3779	16.0
Cech et al. [Čech 2016]	0.2218	33.83
Uricar et al. [Uřičář 2016]	0.2109	32.17

Table 2.2: Landmark localization results on the 300W testing dataset using 68 points. Accuracy is reported as the AUC and the Failure Rate.

Figure 2.8 (top) compares DenseReg+MDM with the results of the latest 300W competition [Sagonas 2016].

We greatly outperform all competitors by a large margin. It should be noted that the participants of the competition did not have any restrictions on the amount of training data employed and some of them are industrial companies (e.g. Fan et al. [Fan 2016]), which further illustrates the effectiveness of our approach. Finally, Table 2.2 reports the area under the curve (AUC) of the CED curves, as well as the failure rate for a maximum RMS error of 0.1. Apart from the accuracy improvement shown by the AUC, we believe that the reported failure rate of 3.67% is remarkable and highlights the robustness of DenseReg.

Method	*AUC*	*Failure Rate (%)*
DenseReg + MDM	**0.5937**	**4.57**
DenseReg	0.4320	8.1
Yang et al. [Yang 2015]	0.5832	4.66
Xiao et al. [Xiao 2015]	0.5800	9.1
Rajamanoharan et al. [Rajamanoharan 2015]	0.5154	9.68
Wu et al. [Wu 2015]	0.4887	15.39
Unicar et al. [Uricár 2015]	0.4059	16.7

Table 2.3: Deformable tracking results against the state-of-the-art on the 300VW testing dataset using 68 points. Accuracy is reported as AUC and the Failure Rate.

2.4.4 Deformable Tracking

For the challenging task of deformable face tracking on lengthy videos, we employ the testing database of the 300VW challenge [Shen 2015, Chrysos 2015] - the only existing benchmark for deformable tracking "in-the-wild". The benchmark consists of 114 videos ($\sim$ $218k$ frames in total) and includes videos captured in totally arbitrary conditions (severe occlusions and extreme illuminations).

The tracking is performed based on sparse landmark points, thus we follow the same strategy as in the case of landmark localization in Sec. 2.4.3.

We compare the output of DenseReg, as well as DenseReg+MDM which was the best performing combination for landmark localization in static images (Sec. 2.4.3), against the participants of the 300VW challenge.

Table 2.3 reports the AUC and Failure Rate measures. DenseReg combined with MDM demonstrates better performance than the winner of the 300VW competition. It should be highlighted that our approach is not fine-tuned for the task-at-hand as opposed to the rest of the methods that were trained on video sequences and most of them make some kind of temporal modelling. Finally, similar to the 300W case, the participants were allowed to use unlimited training data (apart from the provided training seuqences), as opposed to DenseReg (and MDM) that were trained only on the 3148 images mentioned in Sec. 2.4.1. Please refer to the supplementary material for a more detailed presentation of the tracking results.

2.4.5 Monocular Depth Estimation

The fitted template shapes also provide the depth from the image plane. We transfer this information to the visible pixels on the image using the same z-buffering operation used for the deformation-free coordinates (detailed in Sec. 2.2 of the paper). We adopt this as an additional supervision signal: $Z \in [0, 1]$ and add another branch to our network to estimate the depth along with the deformation-free coordinates. To our knowledge, there is no existing results in literature that would allow a quantitative comparison. We are providing example reconstructions using estimated monocular depth fields at Fig.2.9. We observe that this additional branch does not affect the performance of other branches and adds little to the complexity,

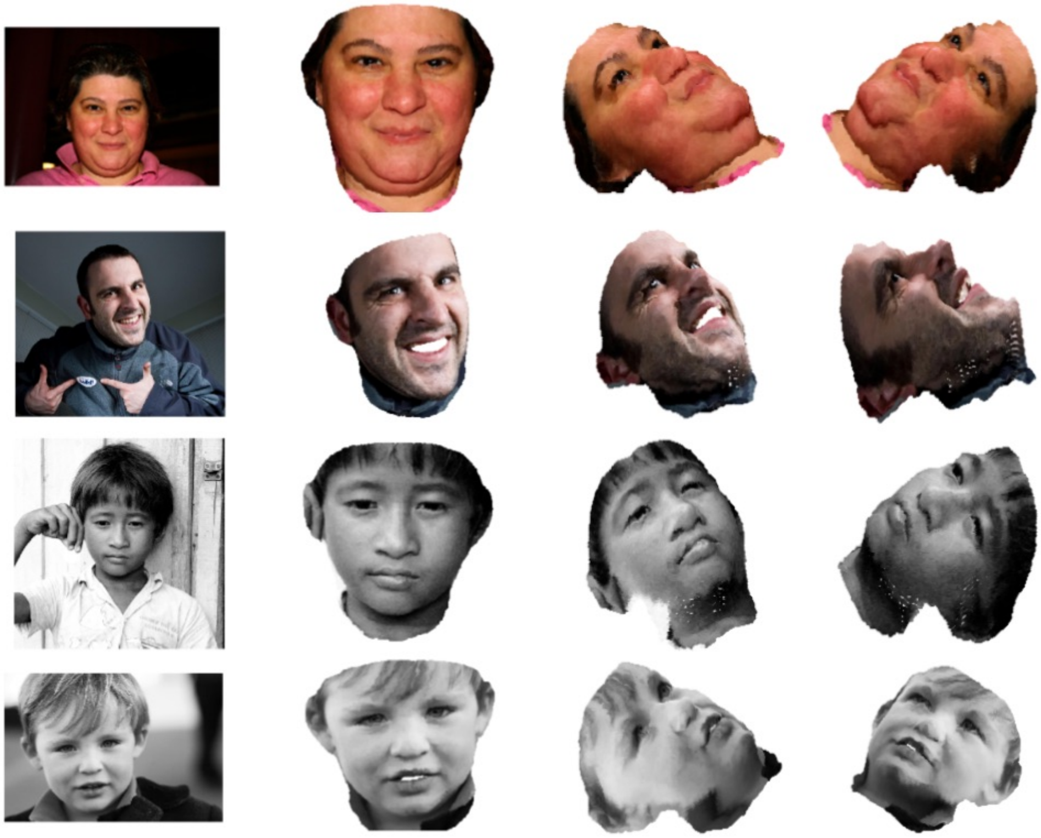


Figure 2.9: Example 3D renderings obtained using estimated depth values.

since it is just a 1x1 convolution layer after the final shared convolutional layer.

2.4.5.1 Ear Shape Regression

We have also performed experiments on the human ear. We employ the 602 images and sparse landmark annotations that were generated in a semi-supervised manner [Zhou 2016c]. Due to the lack of a 3D model of the human ear, we apply Thin Plate Splines to bring the images into dense correspondence and obtain the deformation-free space. We perform landmark localization following the same procedure as in Sec. 2.4.3.

Quantitative results are detailed in the supplementary material, where we compare DenseReg, DenseReg + AAM and DenseReg + MDM with alternative DPM detector based initializations. We observe that DenseReg results are highly accurate and clearly outperforms the DPM based alternative even without a deformable model. Examples for dense human ear correspondence estimated by our system are presented in Fig. 2.10.

The deformation-free space for the ear shape template is visualized in Fig. 2.11. The colouring of the qualitative results that are presented in the are generated using these coordinates. On Table.2.4, we provide failure rates and the Area Under

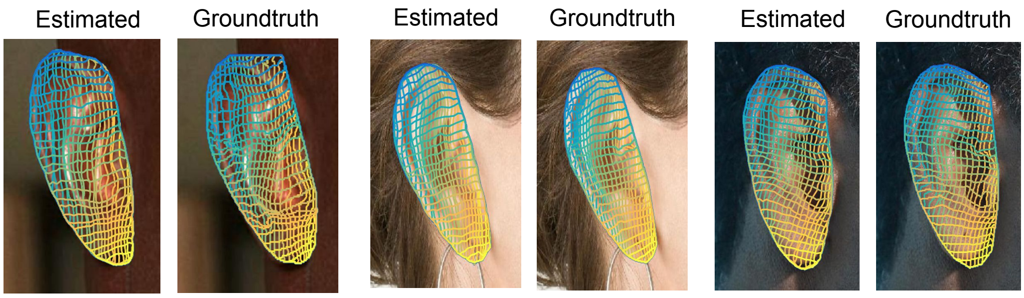


Figure 2.10: Example pairs of deformation-free coordinates of dense landmarks on human ear.

Curve(AUC) measures based on the CED curve of the human ear landmark localization experiment, which were not provided in the paper due to space constraints. Further qualitative examples for regressed and ground-truth deformation-free ear coordinates are provided in Fig. 2.10.

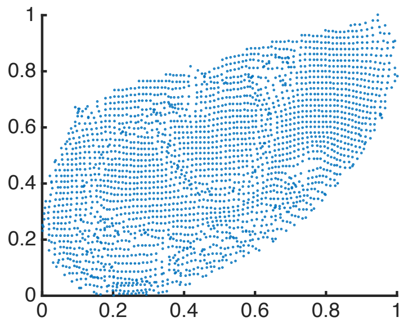


Figure 2.11: Deformation-free space for the template ear shape.

Method	*AUC*	*Failure Rate (%)*
DenseReg + MDM	**0.4842**	**0.98**
DenseReg	0.4150	1.96
DenseReg + AAM	0.4263	0.98
DPM + MDM	0.4160	15.69
DPM + AAM	0.3283	22.55

Table 2.4: Landmark localization results on human ear using 55 points. Accuracy is reported as the Area Under the Curve (AUC) and the Failure Rate of the Cumulative Error Distribution of the normalized RMS point-to-point error.

2.5 Summary

In this chapter, we introduced a fully-convolutional regression approach for establishing dense correspondence fields between objects in natural images and three-dimensional object templates. We demonstrate that the correspondence information can successfully be utilized on problems that can be geometrically represented on the template shape.Furthermore, we unify the problems of dense shape regression and articulated pose of estimation of deformable objects, by proposing the first landmark localization system based on dense shape estimation.

Throughout the chapter, we focused on the human face, where applications are abundant and benchmarks allow a fair comparison. We show that using our dense regression method out-of-the-box outperforms a state-of-the-art semantic segmentation approach for the task of face-part segmentation, while when used as an initialisation for SDMs, we obtain the state-of-the-art results on the challenging 300W landmark localization challenge. We demonstrate the generality of our method by performing experiments on the human ear shapes.

Chapter 3

Quantized Regression and Structured Prediction for Deep Monocular 3D Human Pose Estimation

In this Chapter we focus on the challenging task of 3D human pose estimation from a single monocular image by blending a feed-forward CNN with a graphical model that couples the 3D positions of parts. The CNN populates a volumetric output space that represents the possible positions of 3D human joints and also regresses the estimated displacements between pairs of parts. These constitute the 'unary' and 'pairwise' terms of the energy of a graphical model that resides in a 3D label space and delivers an optimal 3D pose configuration at its output. We show that quantized regression can be used to get a high resolution in the estimation of the pose without increasing the computation/memory requirements.

This work was done in collaboration with Dr. Stefan Kinauer (equal contribution) and published in the Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR 2017).

3.1 Introduction

As reviewed in Sec. 1.2.1.4, prior knowledge about the structure of the 3D human body is commonly incorporated when predicting 3D pose from monocular images, e.g. [Tome 2017]. Two-stage approaches such as [Chen 2016a, Bogo 2016] firstly detect joint positions in 2D and subsequently lift joints into 3D by relying on prior knowledge about the 3D human pose. The advantage of such approaches is that they can exploit large datasets constructed for the prediction of 2D landmarks - the disadvantage is that errors in the 2D stage can propagate to the 3D predictions and can often not be recovered from. Inherently 3D approaches like [Pavlakos 2017] discretize the depth variable and train a CNN to score every possible combination of position and depth with respect to the presence of a joint - one can understand that the CNN learns to use the scale of the joint to guess its depth. This approach delivers results that are largely superior over previous 2-stage approaches.

In directly regressing the pose from the input image, the aforementioned approaches do not *explicitly* impose constraints that exploit the dependencies between the human joints. [Tekin 2016] acknowledge this deficiency of contemporary methods and propose to use a stacked denoising auto-encoder to learn these dependencies implicitly. Other approaches to combining structured prediction with deep learning have recently been successfully pursued in 2D human pose estimation e.g. [Tompson 2014, Yang 2016], while current approaches to incorporating structure in feedforward CNNs for pose estimation rely on cascading, or stacking the outputs of CNNs in 2D [Newell 2016, Wei 2016b], which can become prohibitive when done in 3D, due to the increased memory and computation load. In this work we develop novel techniques that allow us to 'explicitly' capture the dependencies between human joints via an energy function that consists of unary and pairwise terms and thereby pursue this direction in the arguably harder 3D setting.

Our contribution consists in showing that one can combine a volumetric representation with a structured model that imposes constraints between the relative positions of parts. Rather than relying exclusively on a feed-forward architecture, we show that one can append a structured prediction algorithm that optimizes the CNN outputs with respect to the subsequent pose estimation algorithm.

3.2 Methods

We start by formulating our approach in terms of a structured prediction problem, and then provide the details about the individual components of our proposed approach. We represent the pose Φ in terms of the concatenation of the 3D coordinates of N individual parts ϕ_i

$$\Phi = \{\phi_1, \dots, \phi_N\}. \tag{3.1}$$

Given an image I, we score a candidate pose in terms of a graphical model that considers individual properties of parts, as well as properties of some of their pairwise combinations:

$$S_I(\Phi) = \sum_{i=1}^{N} \mathcal{U}_i(\phi_i) + \sum_{i,j \in \mathcal{E}} \mathcal{P}_{i,j}(\phi_i, \phi_j), \tag{3.2}$$

where $\mathcal{U}$ stands for unary and $\mathcal{P}$ for pairwise potentials, and $\mathcal{E}$ is the set of edges used in our graphical model. The unary and pairwise terms are delivered by the CNN, while the structured prediction layer couples the parts through the optimization of Eq. 3.2. If we consider a generic cost function, this can be challenging even for simple cases, let alone for the 3D pose space we are working with. Our main technical contributions aim at making the construction and optimization of Eq. 3.2 tractable while still exploiting the structure of the output space.

3.2.1 Quantized Regression for Depth Estimation

One of the main challenges in constructing a volumetric CNN is that the amount of memory and computation scales linearly in the granularity of the depth quantization, requiring to tradeoff accuracy for speed/memory. The root of the problem is

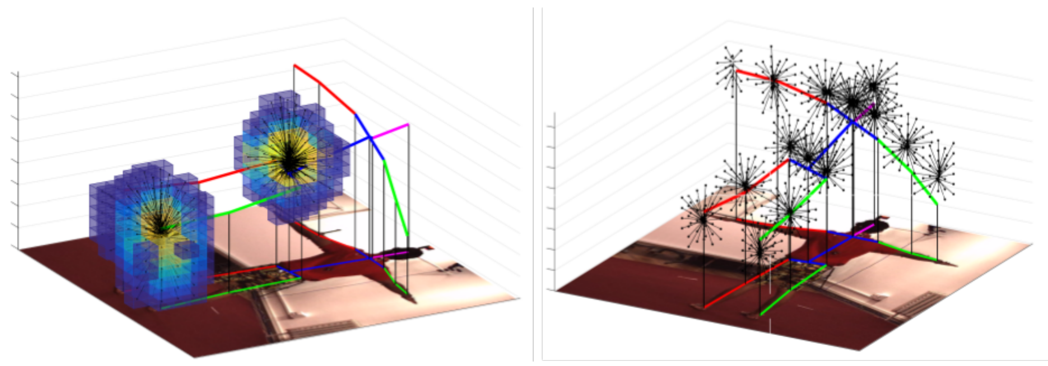


Figure 3.1: Unary 3D coordinates via quantized regression. To efficiently regress the unary 3D coordinates, we use a divide and conquer strategy. We begin by quantizing the 3D space into voxels. We estimate the score of each joint belonging to each of these voxels using a classifier. Finally we regress a residual vector per voxel which indicates the offset between the center of the voxel and the continuous 3D position of each joint. *Left:* Sigmoid function on classified voxels and regressed residual vectors (in black) for two joints. *Right:* Regressed residual vectors for all joints.

that the underlying quantity is continuous, but plain regression-based models may be neither sufficiently accurate, nor expressive enough to capture the uncertainty and multimodality of the depth value caused by depth ambiguity, or occlusion.

Instead, we follow recent successful developments in object detection [Girshick 2015, Ren 2015], dense correspondence estimation as shown in Chap.2, and pose estimation [Papandreou 2017] where a combination of classification and regression is used to attack the image-based regression problem. We use a first classification stage to associate a confidence value with a set of non-overlapping depth intervals, corresponding to a coarse quantization of the depth value. If we have N classes and a depth range of, say D units, the k-th class is associated with a quantized depth of $q_k = k\frac{D}{N}$. This however may be at a very coarse depth resolution. We refine this coarse estimate by combining it with the results of a regression layer that aims at recovering the residual between the ground-truth depth values and their quantized depth estimates.

As shown in Fig. 3.1 this strategy allows us to 'retarget' the voxels to 3D positions that lie closer to the actual part positions, without requiring the exhaustive sampling of the 3D space. In particular a voxel v lying at the k-th depth interval will become associated with a novel 3D position of part i, $p_i^v = k\frac{D}{N} + r_i(v)$, where $r_i(v)$ is the residual regressed by our network for the i-th part type at voxel v.

The value of the associated unary terms, $U_i(p_i^v)$, is obtained in terms of the inner product between a joint-specific weight vector, w_i and a feature vector extracted from the CNN's output at the 2D position associated with voxel v.

3.2.2 Efficient Optimization with Quadratic Pairwise Terms

Having described how the unary terms are constructed in our model, we now turn to the pairwise terms and the resulting optimization problems. The expression for the pairwise term in Eq. 3.2, $\mathcal{P}_{i,j}(\phi_i, \phi_j, I)$ would suggest constructing a six-dimensional function. Instead, as in the Deformable Part Model paradigm [Felzenszwalb 2005], we use a pairwise term that penalizes deviations from a nominal displacement $\mu_{i,j}$:

$$\mathcal{P}_{i,j}(\phi_i, \phi_j, I) = -\sum_{d=1}^{3} c_d(\phi_{d,i} - \phi_{d,j} - \mu_{d,i,j})^2, \tag{3.3}$$

where the c_d parameters allow us to calibrate the importance of the different dimensions. These parameters are forced to be positive, while the expression in Eq. 3.3 corresponds to the log-probability under an axis-aligned Gaussian model, centered at the predicted part position. We note that as in [Chen 2014, Sapp 2010b], $\mu_{i,j}$ is image-dependent, and in our case is the output of a sub-network which is trained end-to-end. This enables us to capture dependencies between parts, where an estimate of their ideal displacement is combined with the local evidence provided by their unary terms.

One important advantage of the pairwise terms is that since they encode the relative position of parts they are often easier to model, since e.g. the distance between human joints is much more predictable than the actual positions of the joints. As such they can simplify the overall problem.

Another crucial advantage of the particular form of the pairwise term is that by virtue of being in the form of a quadratic cost function, it can easily be bounded from above and below using interval arithmetic - in particular, we rely on the 3D Branch-and-Bound algorithm introduced recently in [Kinauer 2016] to efficiently search over optimal combinations of parts in 3D. A brute-force, dynamic programming-type algorithm for solving this task would require a quadratic number of operations, since it would need to compare pairs of points. Our implementation has a low-constant linear complexity for the construction of per-part KD-tree data structures, and logarithmic best-case complexity for the subsequent optimization. In practice optimization requires less than a tenth of a second on a CPU, while further accelerations could be obtained through GPU-based implementations.

3.2.3 Network Connectivity: from star-shaped to loopy graphs

The Branch-and-Bound (BB) algorithm we use for efficient inference only accommodates a star-shaped graph topology. This can be problematic if one wants to model human pose in terms of a tree-structured graph, or introduce loops to capture more constraints. For this we employ master-slave type approximate inference techniques that allow us to use BB for slave problems and coordinate them through a master. In particular we rely on the Alternating Direction Method of Multipliers (ADMM) [Boyd 2011, Martins 2011, Boussaid 2014] which matches the continuous nature of the pose estimation problem [Boussaid 2014]. The approach to subdivide difficult problems into smaller and easier ones has before been seen in [Komodakis 2007]. The authors introduced Dual Decomposition to optimize MRF-type energies, outperforming former state of the art of “tree-reweighted message passing” algorithms. Later works on ADMM like [Boussaid 2014] borrow from developments outlined in [Boyd 2011] to reach convergence in a lower number of iterations. Loopy graphs are subdivided into easier to handle trees and coordinated via a master problem, which turns out to be updating the dual variables.

The method we outline below uses approximate inference to obtain solutions in $\omega(T \log N)$ operations, where T is a low constant in the order of tens, N is the number of voxels, and $\log N$ is the cost of re-solving the slave subproblems. The ω (best-case) notation relates to the (exact) Branch-and-Bound algorithm, which also empirically has typically this performance. Even though the ADMM-based results are now only approximately optimal, the cost function being optimized reflects more accurately the problem structure, which can positively affect accuracy.

We consider the case where the set of graph edges in Eq. 3.2 corresponds to a graph with loops. Denoting by $R \subset 1...K$ the subset of point indices belonging to more than one star graph, our optimization problem can equivalently be rewritten as follows:

$$\max\, S(\Phi) = \sum_{i=1}^{N} S_i(\Phi_i) \quad \text{s.t.} \quad \Phi_i(r) = u(r) \quad \forall r \in R, \tag{3.4}$$

where S_i is a set of loop-free subproblems, defined so that $S(\Phi) = \sum_{i=1}^{N} S_i(\Phi_i)$ for a common solution Φ. The consistency is enforced by the ‘master’, to whom the ‘slave’ subproblems S_i deliver their solutions Φ_i - obtained through Branch-and-Bound. In particular a relaxation to the constraints is updated and used to reset the problem solved by the slaves - at each step the relaxation becomes tighter and at convergence consistency is guaranteed. Dual Decomposition relaxes the constraints in Eq. 3.4 by introducing a Lagrange Multiplier $\lambda_i(r)$ for each agreement constraint. ADMM augments this with a quadratic constraint violation penalty resulting in an *augmented Lagrangian* function:

$$\mathcal{A}(\Phi, u, \lambda) = \sum_{i=1}^{N}(S_i(\Phi_i) + \sum_{r \in R}\langle \lambda_i(r), \Phi_i(r)\rangle) - \sum_{r \in R}((\sum_{i=1}^{N} \lambda_i(r))u(r) - \frac{\rho}{2}\sum_{i=1}^{N}(\Phi_i(r) - u(r))^2) \tag{3.5}$$

where ρ is a positive parameter that controls the intensity of the augmenting penalty. The quadratic term ensures rapid convergence by acting like a regularizer of the solutions found across different iterations. To maximize the augmented Lagrangian, ADMM iteratively performs the following steps:

$$\Phi_i^{t+1} = \arg\max_{\Phi_i} \mathcal{A}(\Phi_i, u^t, \lambda^t) \tag{3.6}$$

$$u^{t+1} = \arg\max_{u} \mathcal{A}(\{\Phi_i^{t+1}\}, u, \lambda^t) \tag{3.7}$$

$$\lambda_i^{t+1}(r) = \lambda_i^t(r) - \rho(\Phi_i^{t+1}(r) - u^{t+1}(r)) \tag{3.8}$$

In words, the slaves efficiently solve their sub-problems and update the master about Φ_i, then the master sets $u^{t+1}(r)$, and the current multipliers $\lambda_i^{t+1}(r)$, and communicates them back to the slaves for the next iteration. Unlike [Boussaid 2014] who used dynamic programming to efficiently solve the slave problems, here we combine ADMM with the Branch-and-Bound algorithm. Interestingly, both of the additional terms contributed by the master problem to the slave problems, $\lambda_i(r)u(r)$, $(\Phi_i(r) - u(r))^2$ can be easily bounded using interval arithmetic, allowing for a straightforward incorporation into the original Branch-and-Bound method. With these changes we have observed similar convergence behavior as the one reported in [Boussaid 2014]; In typically 15-20 (sometimes even less) ADMM iterations the slaves converge to a consistent pose estimate.

3.2.4 Deeply Supervised 2D- and 3D- Learning

We have observed substantial simplifications in the learning procedure by employing Deeply Supervised Network (DSN) [Lee 2015] training. In particular we use loss functions that directly operate on the unary and pairwise terms, before these are integrated through structured prediction. We empirically observed that this substantially accelerates and robustifies learning, by helping the network come up with good 'proposals' to the subsequent combination stage.

As discussed in Sec.3.2.1, the unary coordinates are obtained by adding the quantization and regression signals. Rather than expect this result to be correctly obtained only by back-propagation from the last layer, we also associate a classification and regression problem with each 2D image position.

We associate every pixel with a set of discrete labels corresponding to quantized depth values. For each joint we learn a different classification function; we consider a voxel as being positive if the respective joint is within certain proximity to the 3D location of the ground truth annotation. We train this classifier using the cross-entropy loss. We also regress residual vectors between voxel centers and ground truth joints using an L1 loss which is only active when a voxel is close enough to 3D landmarks.

For the pairwise terms, we regress vectors that point from each 3d joint to others. Similar to the unary coordinates, we regress these quantities in a fully-convolutional manner. The smooth L1 loss for the pairwise offsets between a specific joint and the rest of the joints is only active on pixels within certain proximity to the specific joint.

3.2.5 Training with a Structured Loss Function

Having outlined our cost function and our optimization algorithm, we now turn to parameter estimation. Our graphical model is defined in Eq. 3.2, and the pairwise terms are described in Eq. 3.3. As outlined in the preceding sub-sections, our network generates the unary terms $U_i(p_i^v)$, the nominal displacements $\mu_{i,j}$ and the 3D coordinates ϕ_i. In this section we describe training of all these parameters, as well as the calibration parameters c in Eq. 3.3, using a structured loss function [Joachims 2009,Pepik 2015,Boussaid 2014] that reflects the geometric nature of the problem we want to address. Once our loss function is defined, back-propagation can be used to update all of the underlying network parameters.

While authors in [Zhang 2015a] use an Intersection-over-Union (IoU) based structured loss for the task of detection, given that in this setup we have access to continuous ground truth values that naturally capture the underlying geometry of the problem, we opt for simplicity and use a more straightforward structured loss function.

Given that Φ denotes the 3D coordinates for a candidate configuration of parts (Eq. 3.1), and $\hat{\Phi}$ denotes the groundtruth 3D coordinates, we use the Mean Euclidean Distance, $\Delta(\hat{\Phi}, \Phi) = \frac{1}{P}\sum_{p=1}^{P} \|\phi_p - \hat{\phi}_p\|_2$ as a loss for our learning task, penalizing the 3D displacement of our estimated landmarks from their ground truth positions. As in standard structured output prediction, we use this loss to induce a set of constraints in pose space:

$$S(\hat{\Phi}) > S(\Phi) + \Delta(\hat{\Phi}, \Phi) \;\; \forall \Phi, \tag{3.9}$$

requiring that the score of the ground truth configuration should be greater than the score of any other configuration by a margin depending on how far the particular configuration is from the ground truth.

Since this cannot hold in general, we introduce slack variables ξ: $\xi(\Phi) = max(S(\Phi) + \Delta(\hat{\Phi}, \Phi) - S(\hat{\Phi}), 0)$. Thus, the slack variables represent the violations of the constraints in Eq. 3.9, and our goal here is to learn the model parameters that minimize the slack variables.

Standard training of structural SVMs [Joachims 2009, Pepik 2015, Boussaid 2014] typically finds the most violated configuration given by $\Phi^* = \text{argmax}_\Phi(S(\Phi) + \Delta(\hat{\Phi}, \Phi) - S(\hat{\Phi}))$ and tries to reduce the violation of this configuration by updating the model parameters appropriately via the cutting-planes or Franke-Wolfe algorithm. In this work we use the standard stochastic gradient algorithm to minimize these slack variables. We do so by first finding K most violated configurations for each input sample (K is a hyper-parameter which affects the convergence speed; we set $K = 20$ based on experiments on a validation set). We then compute the subgradients of the model parameters with respect to each of these violated constraints and back-propagate them through the network.

3.3 Experimental Evaluation

Network Architecture

In our experiments we use a fully-convolutional 151 layer ResNet, with weights initialised from a model pre-trained on MPII for 2D body pose estimation [Insafutdinov 2016]. Both 3D and 2D branches of our network are implemented as single-level convolution layers branching from the last layer of the ResNet. The input images to the system are cropped and rescaled to a fixed size of 320x320; the downsampling factor of our network is 16, leading to a cube of 20x20x20 dimensions for 3D unary detection and residual regression branches and 20x20 spatial dimensions for the 2D branches.

Dataset

We use the largest available 3D human pose dataset Human3.6M [Ionescu 2014b] to train and evaluate our approach. The dataset consists of 3.6 million video frames of daily life activities performed by actors whose 3D joint locations are recorded by motion capture systems. Following the recent works in the literature, we have used frames from subjects S1, S5, S6, S7 and S8 for training and S9 and S11 for testing. We have used frames from all 4 cameras and all 15 actions in our training and testing in an action-agnostic manner. We have sub-sampled the videos at 10 frames per second. Several videos that suffer from *drift* of the groundtruth joints are removed from the dataset.

Due to the projective geometry, it is not possible to obtain "groundtruth data-cubes" from 3D poses. In particular, we cannot assume 3D points project to 2D points according to an orthogonal projection model. To cope with this issue, [Pavlakos 2017] create a data-cube using image coordinates for x and y dimensions and real-world coordinates for the z dimension (distances relative to the root node). At test time the depth of the root node and the intrinsic camera parameters are used to obtain 3D pose estimates.

Unlike their approach, which requires knowledge of the root node's z-coordinate at test time, we estimate z coordinates such that the ratio of standard deviations of real-world and projected coordinates in x, y dimensions is preserved in the z dimension. This approximation naturally introduces some reconstruction error, but leads to a system that estimates pose up-to a similarity transform agnostic to the distance of the person to the camera and the intrinsic camera parameters.

Joint Training with 2D Pose

Our network is initialized with ResNet parameters obtained by training for 2D joint localization on the MPII dataset, but we observe that including samples from MPII as training samples increases performance - apparently not doing so results in the network forgetting about 2D joint localization. As in [Sun 2018] we modify the labelled joints of the Human3.6m dataset in order to be able to utilize the 2D data. In particular we include a joint of "thorax" between shoulders that is connected to the "neck" and discarding "chin" and "abdomen" joints. The resulting skeleton structure is identical to the one of MPII. We have verified that two identical networks trained with baseline and MPII-type label structures lead to equivalent evaluation scores, thus it is fair to compare to existing methods. The active losses for an MPII sample are 2D detection and X and Y pairwise offset values, while the 3D position estimates are ignored.

Results

Since our groundtruth comes in the form of projected coordinates, we can obtain the 3D pose only up-to a similarity transform. We report "reconstruction error", which is measured as the mean euclidean distance to the ground truth, after applying Procrustes analysis.

	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases
UNARY alone	49.69	49.45	47.77	50.69	54.80	57.35	43.76	44.11
center star	49.41	49.26	47.35	49.93	50.97	56.12	43.62	**43.43**
stick figure	49.13	49.19	47.15	49.70	50.50	55.57	43.53	43.59
extended stick figure	49.16	49.07	47.35	49.82	50.67	55.45	43.60	43.57
2-hop	**48.89**	**48.75**	**47.07**	**49.40**	**49.82**	**55.31**	**43.30**	43.47
	Sitting	Sit. Down	Smoking	Waiting	Walk Dog	Walking	Walk Tog.	Average
UNARY alone	65.39	95.76	53.53	46.27	51.53	41.59	49.52	53.48
center star	61.50	**78.09**	52.51	45.88	50.63	41.08	49.41	51.42
stick figure	60.14	79.46	51.52	45.74	50.59	40.73	49.33	51.12
extended stick figure	**59.94**	78.51	**51.42**	46.01	50.39	40.89	49.32	51.08
2-hop	60.48	78.20	51.69	**45.63**	**50.16**	**40.74**	**49.17**	**50.87**

Table 3.1: Comparison of average reconstruction errors for different graph topologies.

We experimented with a number of graph topologies and notice that performance depends on the graph structure: **center star** describes the graph topology where all joints are connected to one central root node at the human's torso. It performs better than "unary only", indicating that the body center "knows" something about the other body parts.

stick figure is a graph that directly corresponds to the human skeleton, i.e. the wrist is connected to the elbow, the elbow is connected to the shoulder, and so on. Clearly the shoulder knows better where the elbow has to be than the root node in the torso. This structure clearly performs better than "center star".

extended stick figure is an extension to "stick figure", containing all its edges plus additional connections between the elbows of left and right arm, left and right knee, head to shoulders and torso to knees. This shows that additional loops boost performance, stabilizing against outliers or false evidence.

2-hop follows the human skeleton like "stick figure" and adds connections from every joint to its indirect (2-)neighbours in the skeleton. This connects, for example, hand with shoulder and ankle to hips and left to right knee. "2-hop" performs best, helping to resolve occlusions and improving accuracy.

Our experiments, reported in Table 3.1 clearly indicate that the 2-hop graph topology outperforms all of the other structures that we experimented with. This indeed justifies using approximate inference (ADMM), since these results require employing a loopy graph.

In Table.3.2 we compare the performance of our method to existing methods. Our results indicate that (a) our quantization + regression-based unary network already delivers excellent results, at the level of the current state of the art. (b) Structured prediction yields an additional, quite substantial boost.

We note that there are some methods that only use a single camera or only S-11

	Average error
Yasin et al. [Yasin 2016]	108.3
Rogez et al. [Rogez 2016]	88.1
Tome et al. [Tome 2017]	70.7
Pavlakos et al. [Pavlakos 2017] [1]	53.2
(Ours)Unary	53.48
(Ours)ADMM	50.87

Table 3.2: A comparison of our approach to methods that report reconstruction error in literature.

	cam1	**cam2**	**cam3**	**cam4**	**Average**
S-9	55.62	51.24	56.10	55.22	54.54
S-11	51.14	42.86	47.83	41.90	45.91
Average	53.72	47.64	52.59	49.57	

Table 3.3: Reconstruction errors for videos for specific cameras and test subjects in the Human 3.6M dataset.

frames as test samples and the rest of the videos for training. In order to compare our approach to such works, we present our results per camera and per subject in Tab.3.3. Our results show that we are also outperforming the very recent work of [Sun 2018], who uses only S-11 as test set and obtains 48.3, which is inferior with respect to our S-11 result(45.91), even though we have not used S-9 for training.

We provide qualitative results in Fig.3.2, demonstrating cases where the ADMM inference clearly increases the pose estimation performance. Figure 3.3 shows some example images from the LSP dataset [Johnson 2010] in the left column, augmented with the inferred body skeleton. The other three columns illustrate the plausible 3D structure as inferred by our approach.

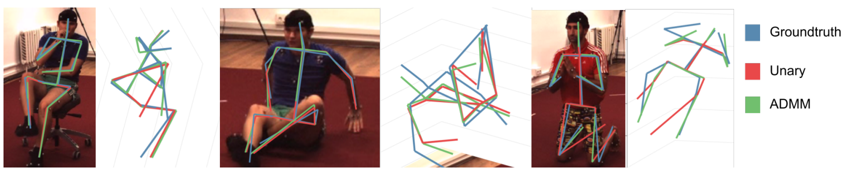


Figure 3.2: Example pose estimates by ADMM inference: Blue indicates the ground truth pose, whereas red and green is the solution obtained from “unaries alone” and ADMM respectively.

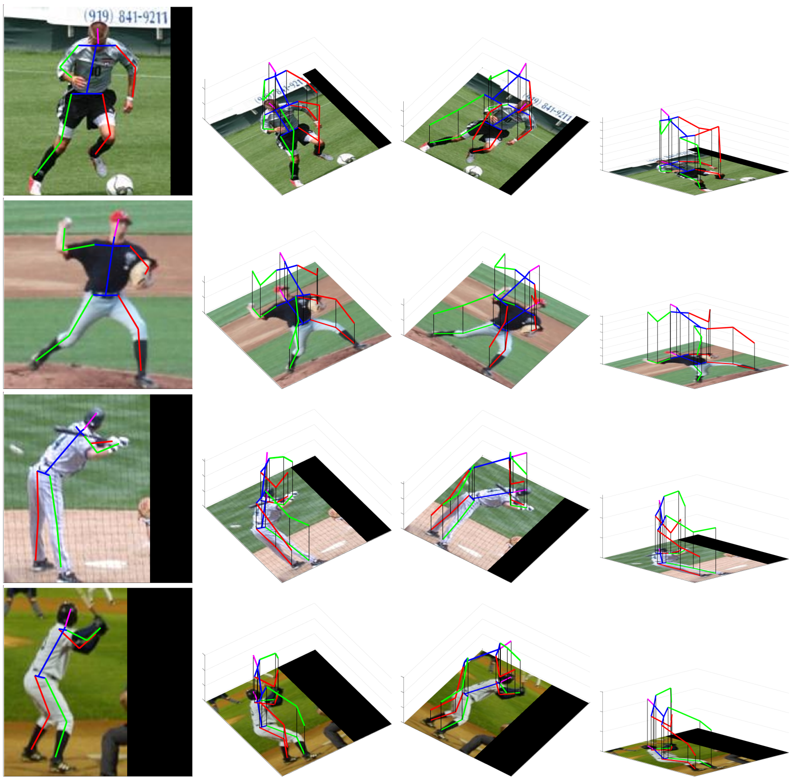


Figure 3.3: Monocular 3D pose estimation results on LSP dataset.

3.4 Summary

In this work we have introduced an efficient method for 3D human pose estimation from 2D images. We have shown that quantized regression can effectively provide 3D volumetric unaries. We report state-of-the-art 3D human pose estimation results, augmenting the functionality of existing deep learning networks by adding a final layer that optimizes an energy function with variables in three dimensions.

DensePose: Dense Human Pose Estimation In The Wild

In this chapter we focus on establishing dense correspondences between an RGB image and a surface-based representation of the human body. We refer to this task as dense human pose estimation. We first gather dense correspondences for 50K persons appearing in the COCO dataset by introducing an efficient annotation pipeline. We then use our dataset to train CNN-based systems that deliver dense correspondence 'in the wild', namely in the presence of background, occlusions and scale variations. We improve our training set's effectiveness by training an 'inpainting' network that can fill in missing ground truth values, and report clear improvements with respect to the best results that would be achievable in the past. We experiment with fully-convolutional networks and region-based models and observe a superiority of the latter; we further improve accuracy through cascading, obtaining a system that delivers highly-accurate results in real time.

This work was published and orally presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. Details on the organized challenges and demonstration videos are provided on the project page `http://densepose.org`.

4.1 Introduction

We introduce the *DensePose* system for the challenging task of establishing dense correspondences between images and a 3D template of the human body. Our work is close in spirit to DenseReg framework, introduced in Chap. 2. DenseReg, supervised by 3DMM supervision mainly focused on faces, and evaluated their results on datasets with moderate pose variability. Here, however, we are facing new challenges, due to the higher complexity and flexibility of the human body, as well as the larger variation in poses. We address these challenges by designing appropriate architectures, as described in Sec. 4.3, which yield substantial improvements over a DenseReg-type fully convolutional architecture. By combining our approach with the recent Mask-RCNN system of [He 2017] we show that a discriminatively trained model can recover highly-accurate correspondence fields for complex scenes involving tens of persons with real-time speed: on a GTX 1080 GPU our system operates at 20-26 frames per second for a 240×320 image or 4-5 frames per second for a 800×1100 image.

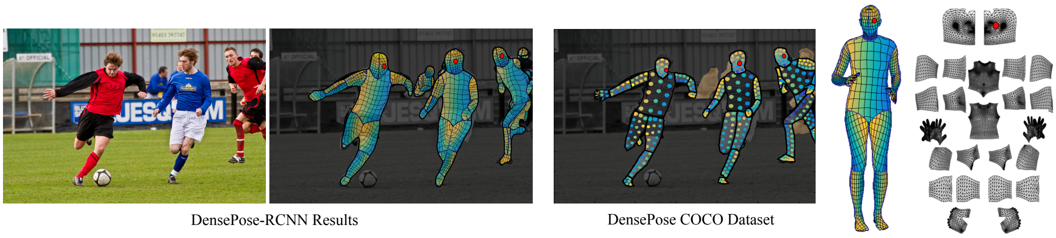


Figure 4.1: Dense pose estimation aims at mapping all human pixels of an RGB image to the 3D surface of the human body. We introduce DensePose-COCO, a large-scale ground-truth dataset with image-to-surface correspondences manually annotated on 50K COCO images and train DensePose-RCNN, to densely regress part-specific UV coordinates within every human region at multiple frames per second. *Left:* The image and the regressed correspondence by DensePose-RCNN, *Middle:* DensePose COCO Dataset annotations, *Right:* Partitioning and UV parametrization of the body surface.

The task of establishing dense correspondences from an image to a surface-based human body model has been addressed mostly in the setting where a depth sensor is available, as in the Vitruvian manifold of [Taylor 2012], metric regression forests [Pons-Moll 2015b], or the more recent dense point cloud correspondence of [Wei 2016a]. By contrast, in our case, we consider a single RGB image as input, based on which we establish a correspondence between surface points and image pixels.

The analysis of people in images and videos is often based on human parts, a coarsened version of image-to-surface correspondence, or landmark detectors, a sparse description of the human body via keypoints such as the elbows, shoulders and ankles, etc. Our approach can be understood as the next step in the line of works on extending the standard 2D and 3D pose estimation for humans.

Our contributions can be summarized in three points. Firstly, as described in Sec. 4.2, we introduce the first manually-collected ground truth dataset for the task, by gathering dense correspondences between the SMPL model [Loper 2015] and persons appearing in the COCO dataset. This is accomplished through a novel annotation pipeline that exploits 3D surface information during annotation.

Secondly, as described in Sec. 4.3, we use the resulting dataset to train CNN-based systems that deliver dense correspondence 'in the wild', by regressing body surface coordinates at any image pixel. We experiment with both fully-convolutional architectures, relying on Deeplab [Chen 2018b], and also with region-based systems, relying on Mask-RCNN [He 2017], observing a superiority of region-based models over fully-convolutional networks. We also consider cascading variants of our approach, yielding further improvements over existing architectures.

Thirdly, we explore different ways of exploiting our constructed ground truth information. Our supervision signal is defined over a randomly chosen subset of image pixels per training sample. We use these sparse correspondences to train a

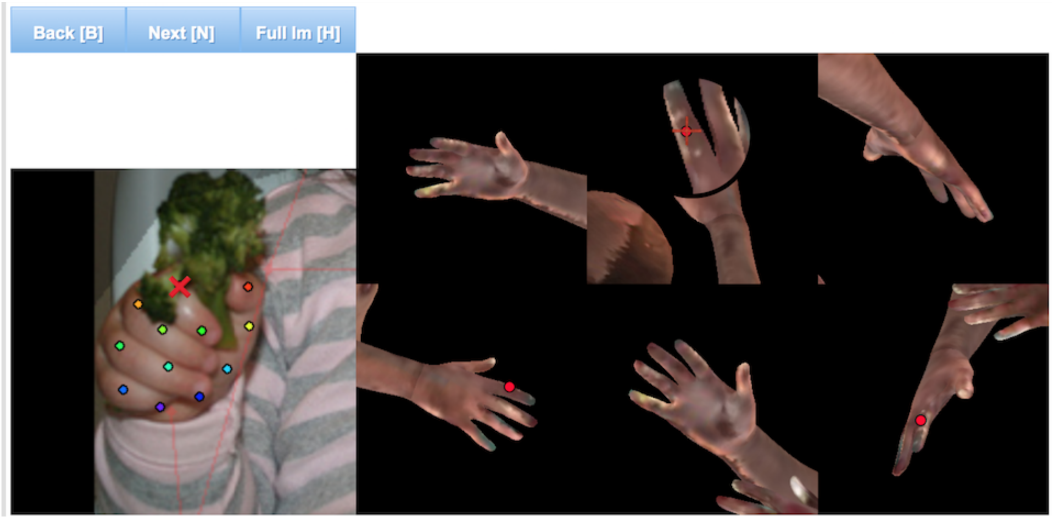


Figure 4.2: The user interface for collecting per-part correspondence annotations: We provide the annotators six pre-rendered views of a body part such that the whole part-surface is visible. Once the target point is annotated, the point is displayed on all rendered images simultaneously.

'teacher' network that can 'inpaint' the supervision signal in the rest of the image domain. Using this inpainted signal results in clearly better performance when compared to either sparse points, or any other existing dataset, as shown experimentally in Sec. 4.4.

Our experiments indicate that dense human pose estimation is to a large extent feasible, but still has space for improvement. We conclude our paper with some qualitative results and directions that show the potential of the method. We will make code and data publicly available from our project's webpage, `http://densepose.org`.

4.2 COCO-DensePose Dataset

Gathering rich, high-quality training sets has been a catalyst for progress in the classification [Deng 2009], detection and segmentation [Everingham 2015, Lin 2014] tasks. There currently exists no manually collected ground-truth for dense human pose estimation for real images. The works of [Lassner 2017b] and [Varol 2017] can be used as surrogates, but as we show in Sec. 4.4 provide worse supervision.

In this Section we introduce our COCO-DensePose dataset, alongside with evaluation measures that allow us to quantify progress in the task in Sec. 4.4. We have gathered annotations for 50K humans, collecting more then 5 million manually annotated correspondences.

We start with a presentation of our annotation pipeline, since this required several design choices that may be more generally useful for 3D annotation. We then turn to an analysis of the accuracy of the gathered ground-truth, alongside

Figure 4.3: Visualization of annotations: Image (left), U (middle) and V (right) values for the collected points.

with the resulting performance measures used to assess the different methods.

4.2.1 Annotation System

In this work, we involve human annotators to establish dense correspondences from 2D images to surface-based representations of the human body. If done naively, this would require 'hunting vertices' for every 2D image point, by manipulating a surface through rotations - which can be frustratingly inefficient. Instead, we construct an annotation pipeline through which we can efficiently gather annotations for image-to-surface correspondence.

As shown in Fig. 2.3, in the first stage we ask annotators to delineate regions corresponding to visible, semantically defined body parts. These include Head, Torso, Lower/Upper Arms, Lower/Upper Legs, Hands and Feet. In order to use simplify the UV parametrization we design the parts to be isomorphic to a plane, partitioning the limbs and torso into lower-upper and frontal-back parts.

For *head*, *hands* and *feet*, we use the manually obtained UV fields provided in the SMPL model [Loper 2015]. For the rest of the parts we obtain the unwrapping via multi-dimensional scaling applied to pairwise geodesic distances. The UV fields for the resulting 24 parts are visualized in Fig. 4.1 (right).

We instruct the annotators to estimate the body part behind the clothes, so that for instance wearing a large skirt would not complicate the subsequent annotation of correspondences. In the second stage we sample every part region with a set of roughly equidistant points obtained via k-means and request the annotators to bring these points in correspondence with the surface. The number of sampled points varies based on the size of the part and the maximum number of sampled points per part is 14. In order to simplify this task we 'unfold' the part surface by providing six pre-rendered views of the same body part and allow the user to place landmarks on any of them Fig. 4.2. This allows the annotator to choose the most convenient point of view by selecting one among six options instead of manually rotating the surface.

As the user indicates a point on any of the rendered part views, its surface coordinates are used to simultaneously show its position on the remaining views – this gives a global overview of the correspondence. The image points are presented to the annotator in a horizontal/vertical succession, which makes it easier to deliver geometrically consistent annotations by avoiding self-crossings of the surface. This two-stage annotation process has allowed us to very efficiently gather highly accurate correspondences. If we quantify the complexity of the annotation task in terms of the time it takes to complete it, we have seen that the part segmentation and correspondence annotation tasks take approximately the same time, which is surprising given the more challenging nature of the latter task. Visualizations of the collected annotations are provided in Fig. 4.3, where the partitioning of the surface and U, V coordinates are shown in Fig. 4.1.

4.2.2 Accuracy of human annotators

We assess human annotator with respect to a gold-standard measure of performance. Typically in pose estimation one asks multiple annotators to label the same landmark, which is then used to assess the variance in position, e.g. [Lin 2014, Ronchi 2017]. In our case, we can render images where we have access to the true mesh coordinates used to render a pixel. We thereby directly compare the true position used during rendering and the one estimated by annotators, rather than first estimating a 'consensus' landmark location among multiple human annotators.

In particular, we provide annotators with synthetic images generated through the exact same surface model as the one we use in our ground-truth annotation, exploiting the rendering system and textures of [Varol 2017]. We then ask annotators to bring the synthesized images into correspondence with the surface using our annotation tool, and for every image k estimate the geodesic distance $d_{i,k}$ between the correct surface point, i and the point estimated by human annotators $\hat{i}_k$:

$$d_{i,k} = g(i, \hat{i}_k), \tag{4.1}$$

where $g(\cdot,\cdot)$ measures the geodesic distance between two surface points.

For any image k, we annotate and estimate the error only on a randomly sampled set of surface points $\mathcal{S}_k$ and interpolate the errors on the remainder of the surface. Finally, we average the errors across all K examples used to assess annotator performance.

As shown in Fig. 4.4 the annotation errors are substantially smaller on small surface parts with distinctive features that could help localization (face, hands, feet), while on larger uniform areas that are typically covered by clothes (torso, back, hips) the annotator errors can get larger.

4.2.3 Evaluation Measures

We consider two different ways of summarizing correspondence accuracy over the whole human body, including pointwise and per-instance evaluation.

Pointwise evaluation. This approach evaluates correspondence accuracy over the whole image domain through the Ratio of Correct Point (RCP) correspondences, where a correspondence is declared correct if the geodesic distance is below a certain threshold. As the threshold t varies, we obtain a curve $f(t)$, whose area provides us with a scalar summary of the correspondence accuracy. For any given image we have a varying set of points coming with ground-truth signals. We summarize performance on the ensemble of such points, gathered across images. We evaluate the area under the curve (AUC), $\text{AUC}_a = \frac{1}{a}\int_0^a f(t)dt$, for two different values of $a = 10\text{cm}, 30\text{cm}$ yielding AUC_{10} and AUC_{30} respectively, where AUC_{10} is understood as being an accuracy measure for more refined correspondence. This performance measure is easily applicable to both single- and multi-person scenarios and can deliver directly comparable values. In Fig. 4.5, we provide the per-part pointwise evaluation of the human annotator performance on synthetic data, which can be seen as an upper bound for the performance of our systems.

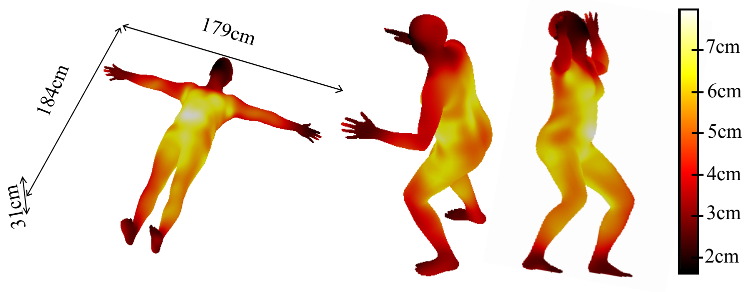


Figure 4.4: Average human annotation error as a function of surface position.

Per-instance evaluation. Inspired by the object keypoint similarity (OKS) measure used for pose evaluation on the COCO dataset [Lin 2014, Ronchi 2017], we introduce *geodesic point similarity (GPS)* as a correspondence matching score:

$$\mathrm{GPS}_j = \frac{1}{|P_j|} \sum_{p \in P_j} \exp\left(\frac{-g(i_p, \hat{i}_p)^2}{2\kappa^2} \right), \tag{4.2}$$

where P_j is the set of ground truth points annotated on person instance j, i_p is the vertex estimated by a model at point p, $\hat{i}_p$ is the ground truth vertex p and κ is a normalizing parameter. We set κ=0.255 so that a single point has a GPS value of 0.5 if its geodesic distance from the ground truth equals the average half-size of a body segment, corresponding to approximately 30 cm. Intuitively, this means that a score of GPS $\approx$ 0.5 can be achieved by a perfect part segmentation model, while going above that also requires a more precise localization of a point on the surface.

Once the matching is performed, we follow the COCO challenge protocol [Lin 2014, Ronchi 2017] and evaluate Average Precision (AP) and Average Recall (AR) at a number of GPS thresholds ranging from 0.5 to 0.95, which corresponds to the range of geodesic distances between 0 and 30 cm. We use the same range of distances to perform both per-instance and per-point evaluation.

4.3 Learning Dense Human Pose Estimation

We now turn to the task of training a deep network that predicts dense correspondences between image pixels and surface points. this work, we introduce improved architectures by combining the DenseReg approach (Chap. 2) with the Mask-RCNN architecture [He 2017], yielding our 'DensePose-RCNN' system. We develop cascaded extensions of DensePose-RCNN that further improve accuracy and describe a training-based interpolation method that allows us to turn a sparse supervision signal into a denser and more effective variant.

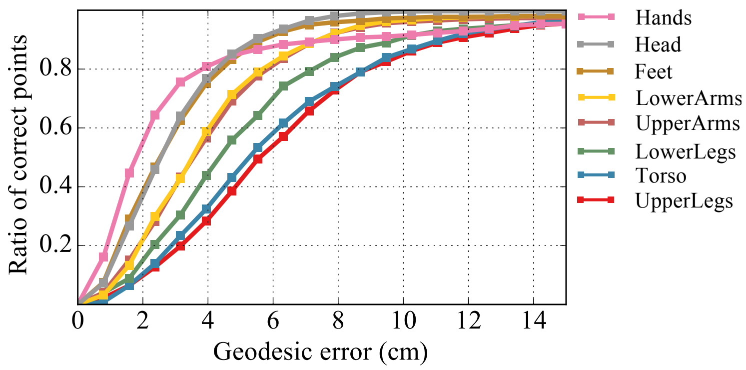


Figure 4.5: Human annotation error distribution within different body parts.

4.3.1 Fully-convolutional dense pose regression

The simplest architecture choice consists in using a fully convolutional network (FCN) that combines a classification and a regression task, similar to DenseReg. In a first step, we classify a pixel as belonging to either background, or one among several region parts which provide a coarse estimate of surface coordinates. This amounts to a labelling task that is trained using a standard cross-entropy loss. In a second step, a regression system indicates the exact coordinates of the pixel within the part. Since the human body has a complicated structure, we break it into multiple independent pieces and parameterize each piece using a local two-dimensional coordinate system, that identifies the position of any node on this surface part.

Intuitively, we can say that we first use appearance to make a coarse estimate of where the pixel belongs to and then align it to the exact position through some small-scale correction. Concretely, coordinate regression at an image position i can be formulated as follows:

$$c^* = \arg\max_c P(c|i), \quad [U, V] = R^{c^*}(i) \tag{4.3}$$

where in the first stage we assign position i to the body part c^* that has highest posterior probability, as calculated by the classification branch, and in the second stage we use the regressor R^{c^*} that places the point i in the continuous U, V coordinates parametrization of part c^*. In our case, c can take 25 values (one is background), meaning that P_x is a 25-way classification unit, and we train 24 regression functions R^c, each of which provides 2D coordinates within its respective part c. While training, we use a cross-entropy loss for the part classification and a smooth L_1 loss for training each regressor. The regression loss is only taken into account for a part if the pixel is within the specific part.

4.3.2 Region-based Dense Pose Regression

Using an FCN makes the system particularly easy to train, but loads the same deep network with too many tasks, including part segmentation and pixel localization, while at the same time requiring scale-invariance which becomes challenging for humans in COCO. Here we adopt the region-based approach of [Ren 2015, He 2017], which consists in a cascade of proposing regions-of-interest (ROI), extracting region-adapted features through ROI pooling [He 2014, He 2017] and feeding the resulting features into a region-specific branch. Such architectures decompose the complexity of the task into controllable modules and implement a scale-selection mechanism through ROI-pooling. At the same time, they can also be trained jointly in an end-to-end manner [Ren 2015].

We adopt the settings introduced in [He 2017], involving the construction of Feature Pyramid Network [Lin 2017] features, and ROI-Align pooling, which have been shown to be important for tasks that require spatial accuracy. We adapt this

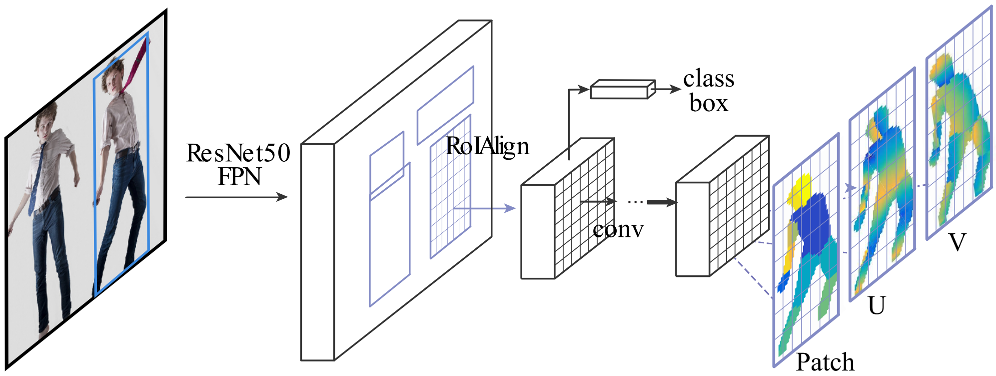


Figure 4.6: DensePose-RCNN architecture: we use a cascade of region proposal generation and feature pooling, followed by a fully-convolutional network that densely predicts discrete part labels and continuous surface coordinates.

architecture to our task, so as to obtain dense part labels and coordinates within each of the selected regions.

As shown in Fig. 4.6, we introduce a fully-convolutional network on top of ROI-pooling that is entirely devoted to these two tasks, generating a classification and a regression head that provide the part assignment and part coordinate predictions, as in DenseReg. For simplicity, we use the exact same architecture used in the keypoint branch of Mask-RCNN, consisting of a stack of 8 alternating 3×3 fully convolutional and ReLU layers with 512 channels. At the top of this branch we have the same classification and regression losses as in the FCN baseline, but we now use a supervision signal that is cropped within the proposed region.

During inference, our system operates at 25fps on 320x240 images and 4-5fps on 800x1100 images using a GTX1080 graphics card.

4.3.3 Multi-task cascaded architectures

Inspired by the success of recent pose estimation models based on iterative refinement [Wei 2016b, Newell 2016] we experiment with cascaded architectures. Cascading can improve performance both by providing context to the following stages, and also through the benefits of deep supervision [Lee 2015].

As shown in Fig. 4.7, we do not confine ourselves to cascading within a single task, but also exploit information from related tasks, such as keypoint estimation and instance segmentation, which have successfully been addressed by the Mask-RCNN architecture [He 2017]. This allows us to exploit task synergies and the complementary merits of different sources of supervision.

4.3.4 Distillation-based ground-truth interpolation

Even though we aim at dense pose estimation at test time, in every training sample we annotate only a sparse subset of the pixels, approximately 100-150 per human. This does not necessarily pose a problem during training, since we can make our

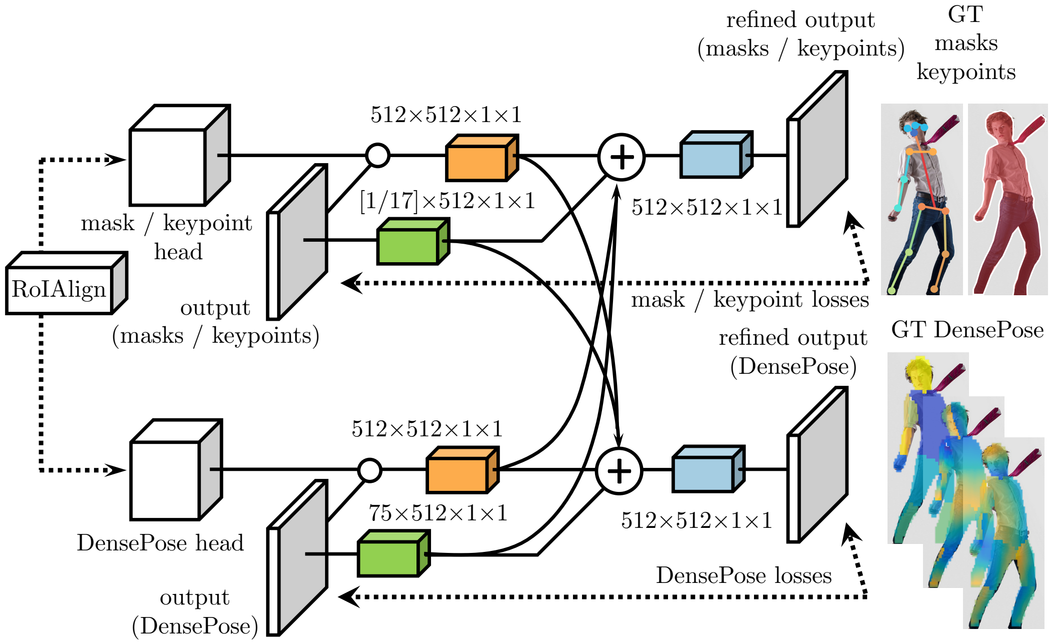


Figure 4.7: Cross-cascading architecture: The output of the RoIAlign module in Fig. 4.6 feeds into the DensePose network as well as auxiliary networks for other tasks (masks, keypoints). Once first-stage predictions are obtained from all tasks, they are combined and then fed into a second-stage refinement unit of each branch.

classification/regression losses oblivious to points where the ground-truth correspondence was not collected, simply by not including them in the summation over the per-pixel losses [Long 2015]. However, we have observed that we obtain substantially better results by "inpainting" the values of the supervision signal on positions that were not originally annotated. For this we adopt a learning-based approach where we firstly train a "teacher" network (depicted in Fig. 4.8) to reconstruct the ground-truth values wherever these are observed, and then deploy it on the full image domain, yielding a dense supervision signal. In particular, we only keep the network's predictions on areas that are labelled as foreground, as indicated by the part masks collected by humans, in order to ignore network errors on background regions.

4.4 Experiments

In all of the following experiments, we assess the methods on a test set of 1.5k images containing 2.3k humans, using as training set of 48K humans. Our test-set coincides with the COCO keypoints-minival partition used by [He 2017] and the training set with the COCO-train partition. We are currently collecting annotations for the remainder of the COCO dataset, which will soon allow us to also have a competition mode evaluation.

Before assessing dense pose estimation 'in the-wild' in Sec. 4.4.3, we start in Sec. 4.4.1 with the more restricted 'Single-Person' setting where we use as inputs images cropped around ground-truth boxes. This factors out the effects of detection performance and provides us with a controlled setting to assess the usefulness of the COCO-DensePose dataset.

4.4.1 Single-Person Dense Pose Estimation

We start in Sec. 4.4.1.1 by comparing the COCO-DensePose dataset to other sources of supervision for dense pose estimation and then in Sec. 4.4.1.2 compare the performance of the model-based system of [Bogo 2016] with our discriminatively-trained system. Clearly the system of [Bogo 2016] was not trained with the same amount of data as our model; this comparison therefore serves primarily to show the merit of our large-scale dataset for discriminative training.

4.4.1.1 Manual supervision versus surrogates

We start by assessing whether COCO-DensePose improves the accuracy of dense pose estimation with respect to the prior semi-automated, or synthetic supervision signals described below.

A semi-automated method is used for the 'Unite the People' (UP) dataset of [Lassner 2017b], where human annotators verified the results of fitting the SMPL 3D deformable model [Loper 2015] to 2D images. However, model fitting often fails in the presence of occlusions, or extreme poses, and is never guaranteed to be entirely successful – for instance, even after rejecting a large fraction of the fitting results, the feet are still often misaligned in [Lassner 2017b]. This both decimates the training set and obfuscates evaluation, since the ground-truth itself may have systematic errors.

Synthetic ground-truth can be established by rendering images using surface-based models [Pishchulin 2011, Pishchulin 2012, Rogez 2016, Ghezelghieh 2016, Chen 2016c, Neverova 2017]. This has recently been applied to human pose in the SURREAL dataset of [Varol 2017], where the SMPL model [Loper 2015] was rendered with the CMU Mocap dataset poses [MoCap 2003]. However, covariate shift can emerge because of the different statistics of rendered and natural images.

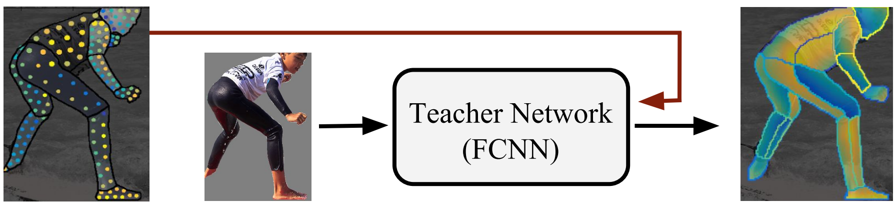
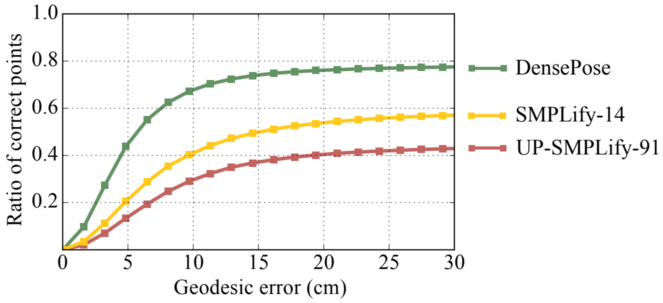


Figure 4.8: We first train a 'teacher network' with our sparse, manually-collected supervision signal, and then use the network to 'inpaint' a dense supervision signal used to train our region-based system.



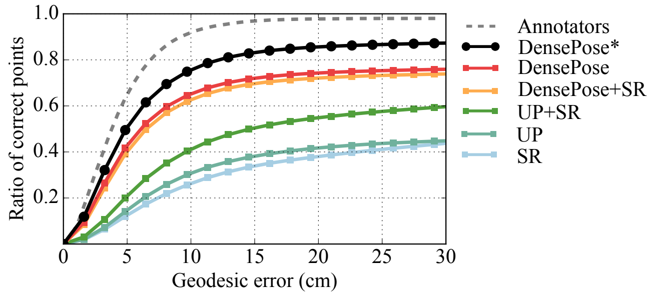
Method	**AUC**$_{10}$	**AUC**$_{30}$
Full-body images		
UP-SMPLify-91	0.155	0.306
SMPLify-14	0.226	0.416
DensePose	0.429	0.630
All images		
SMPLify-14	0.099	0.19
DensePose	0.378	0.614
Human Performance	0.563	0.835

Figure 4.9: Comparison between model-based single-person pose estimation of SMPLify [Bogo 2016] and our FCN-based result, in the absence ('full-body images') and presence ('all images') of occlusions.

Since both of these two methods use the same SMPL surface model as the one we use in our work, we can directly compare results, and also combine datasets. We render our dense coordinates and our dense part labels on the SMPL model for all 8514 images of UP dataset and 60k SURREAL models for comparison.

In Fig. 4.10 we assess the test performance of ResNet-101 FCNs of stride 8 trained with different datasets, using a Deeplab-type architecture. During training we augment samples from all of the datasets with scaling, cropping and rotation. We observe that the surrogate datasets lead to weaker performance, while their combination yields improved results. Still, their performance is substantially lower than the one obtained by training on our DensePose dataset, while combining the DensePose with SURREAL results in a moderate drop in network performance. Based on these results we rely exclusively on the DensePose dataset for training in the remaining experiments, even though domain adaptation could be used in the future [Ganin 2015] to exploit synthetic sources of supervision.

The last line in the table of Fig. 4.10 ('DensePose*') indicates an additional performance boost that we get by using the COCO human segmentation masks in order to replace background intensities with an average intensity during both training and testing and also by evaluating the network at multiple scales and averaging the results. Clearly, the results with other methods are not directly comparable, since we are using additional information to remove background structures. Still,



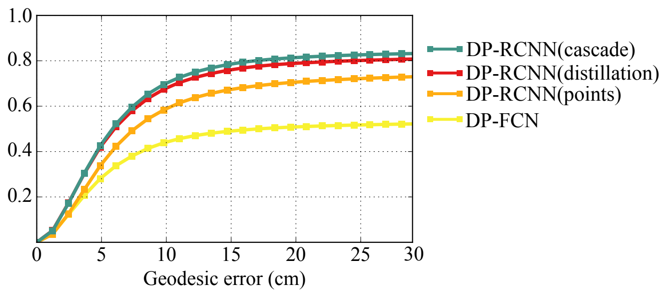
Method	**AUC$_{10}$**	**AUC$_{30}$**
SR	0.124	0.289
UP	0.146	0.319
SR + UP	0.201	0.424
DensePose + SR	0.357	0.592
DensePose	0.378	0.614
DensePose*	0.445	0.711
Human Performance	0.563	0.835

Figure 4.10: Single-person performance for different kinds of supervision signals used for training: DensePose leads to substantially more accurate results than surrogate datasets. DensePose* uses a figure-ground oracle at both training and test time.

the resulting predictions are substantially closer to human performance – we therefore use this as the 'teacher network' to obtain dense supervision for the experiments in Sec. 4.4.2.

4.4.1.2 FCNN- vs Model-based pose estimation

In Fig. 4.9 we compare our method to the SMPLify pipeline of [Bogo 2016], which fits the 3D SMPL model to an image based on a pre-computed set of landmark points. We use the code provided by [Lassner 2017b] with both DeeperCut pose estimation landmark detector [Insafutdinov 2016] for 14-landmark results and with the 91-landmark alternative proposed in [Lassner 2017b]. Note that these landmark detectors were trained on the MPII dataset. Since the whole body is visible in the MPII dataset, for a fair comparison we separately evaluate on images where 16/17 or 17/17 landmarks are visible and on the whole test set. We observe that while being orders of magnitude faster (0.04-0.25" vs 60-200") our bottom-up, feedforward method largely outperforms the iterative, model fitting result. As mentioned above, this difference in accuracy indicates the merit of having at our disposal DensePose-COCO for discriminative training.



Method	$\mathbf{AUC_{10}}$	$\mathbf{AUC_{30}}$	**IoU**
DP-FCN	0.253	0.418	0.66
DP-RCNN (points only)	0.315	0.567	0.75
DP-RCNN (distillations)	0.381	0.645	0.79
DP-RCNN (cascade)	0.390	0.664	0.81
DP*	0.417	0.683	–
Human Performance	0.563	0.835	–

Figure 4.11: Results of multi-person dense correspondence labelling. Here we compare the performance of our proposed DensePose-RCNN system against the fully-convolutional alternative on realistic images from the COCO dataset including multiple persons with high variability in scales, poses and backgrounds.

4.4.2 Multi-Person Dense Pose Estimation

Having established the merit of the DensePose-COCO dataset, we now turn to examining the impact of network architecture on dense pose estimation in-the-wild. In Fig. 4.11 we summarize our experimental findings using the same RCP measure used in Fig. 4.10.

We observe firstly that the FCN-based performance in-the-wild (curve 'DensePose-FCN') is now dramatically lower than that of the DensePose curve in Fig. 4.11. Even though we apply a multi-scale testing strategy that fuses probabilities from multiple runs using input images of different scale [Zhao 2016], the FCN is not sufficiently robust to deal with the variability in object scale.

We then observe in curve 'DensePose-RCNN' a big boost in performance thanks to switching to a region-based system. The networks up to here have been trained using the sparse set of points that have been manually annotated. In curve 'DensePose-RCNN-Distillation' we see that using the dense supervision signal delivered by our DensePose* system on the training set yields a substantial improvement. Finally, in 'DensePose-RCNN-Cascade' we show the performance achieved thanks to the introduction of cascading: Sec. 4.3.3 almost matches the 'DensePose*' curve of Fig. 4.10.

This is a remarkably positive result: as described in Sec. 4.4.1, the 'DensePose*' curve corresponds to a very privileged evaluation, involving (a) cropping objects around their ground-truth boxes and fixing their scale (b) removing background

Figure 4.12: Qualitative evaluation of DensePose-RCNN. *Left:* input, *Right:* DensePose-RCNN estimates. We observe that our system successfully estimates body pose regardless of skirts or dresses, while handling a large variability of scales, poses, and occlusions.

Method	**AP**	**AP_{50}**	**AP_{75}**	**AP_M**	**AP_L**	**AR**	**AR_{50}**	**AP_{75}**	**AR_M**	**AR_L**
DP (ResNet-50)	51.0	83.5	54.2	39.4	53.1	60.1	88.5	64.5	42.0	61.3
DP (ResNet-101)	51.8	83.7	56.3	42.2	53.8	61.1	88.9	66.4	45.3	62.1
Multi-task learning										
DP + masks	51.9	85.5	54.7	39.4	53.9	61.1	89.7	65.5	42.0	62.4
DP + keypoints	52.8	85.6	56.2	42.2	54.7	62.6	89.8	67.7	45.4	63.7
Multi-task learning with cascading										
DP-cascade	51.6	83.9	55.2	41.9	53.4	60.4	88.9	65.3	43.3	61.6
DP + masks	52.8	85.5	56.1	40.3	54.6	62.0	89.7	67.0	42.4	63.3
DP + keypoints	55.8	87.5	61.2	48.4	57.1	63.9	91.0	69.7	50.3	64.8

Table 4.1: Per-instance evaluation of DensePose-RCNN performance on COCO `minival` subset. All multi-task experiments are based on ResNet-50 architecture. DensePose-cascade corresponds to the base architecture with an iterative refinement module with no input from other tasks.

variation from both training and testing, by using ground-truth object masks and (c) ensembling over scales. It can therefore be understood as an upper bound of what we could expect to obtain when operating in-the-wild. We see that our best system is marginally below that level of performance, which clearly reveals the power of the three modifications we introduce, namely region-based processing, inpainting the supervision signal, and cascading.

In Tab. 4.1 we report the AP and AR metrics described in Sec. 4.2 as we change different choices in our architecture. We have conducted experiments using both ResNet-50 and ResNet-101 backbones and observed an only insignificant boost in performance with the larger model (first two rows in Tab. 4.1). The rest of our experiments are therefore based on the ResNet-50-FPN version of DensePose-RCNN. The following two experiments shown in the middle section of Tab. 4.1 indicate the impact on multi-task learning.

Augmenting the network with the mask or keypoint branches yields improvements with any of these two auxiliary tasks. The last section of Tab. 4.1 reports improvements in dense pose estimation obtained through cascading using the network setup from Fig. 4.7. Incorporating additional guidance in particular from the keypoint branch significantly boosts performance.

4.4.3 Qualitative Results

In this section we provide additional qualitative results to further demonstrate the performance of our method. In Fig. 4.12 we show qualitative results generated by our method, where the correspondence is visualized in terms of 'fishnets', namely isocontours of estimated UV coordinates that are superimposed on humans. As these results indicate, our method is able to handle large amounts of occlusion, scale, and pose variation, while also successfully hallucinating the human body behind clothes such as dresses or skirts.

In Fig.4.13 we demonstrate a simple graphics-oriented application, where we map texture RGB intensities taken from [Varol 2017] to estimated UV body coordinates - the whole video is available on our project's website `http://densepose.org`.

4.5 Summary

In this chapter we have tackled the task of dense human pose estimation using discriminative trained models. We have introduced COCO-DensePose, a large-scale dataset of ground-truth image-surface correspondences and developed novel architectures that allow us to recover highly-accurate dense correspondences between images and the body surface in multiple frames per second. We anticipate that this will pave the way both for downstream tasks in augmented reality or graphics, but also help us tackle the general problem of associating images with semantic 3D object representations.

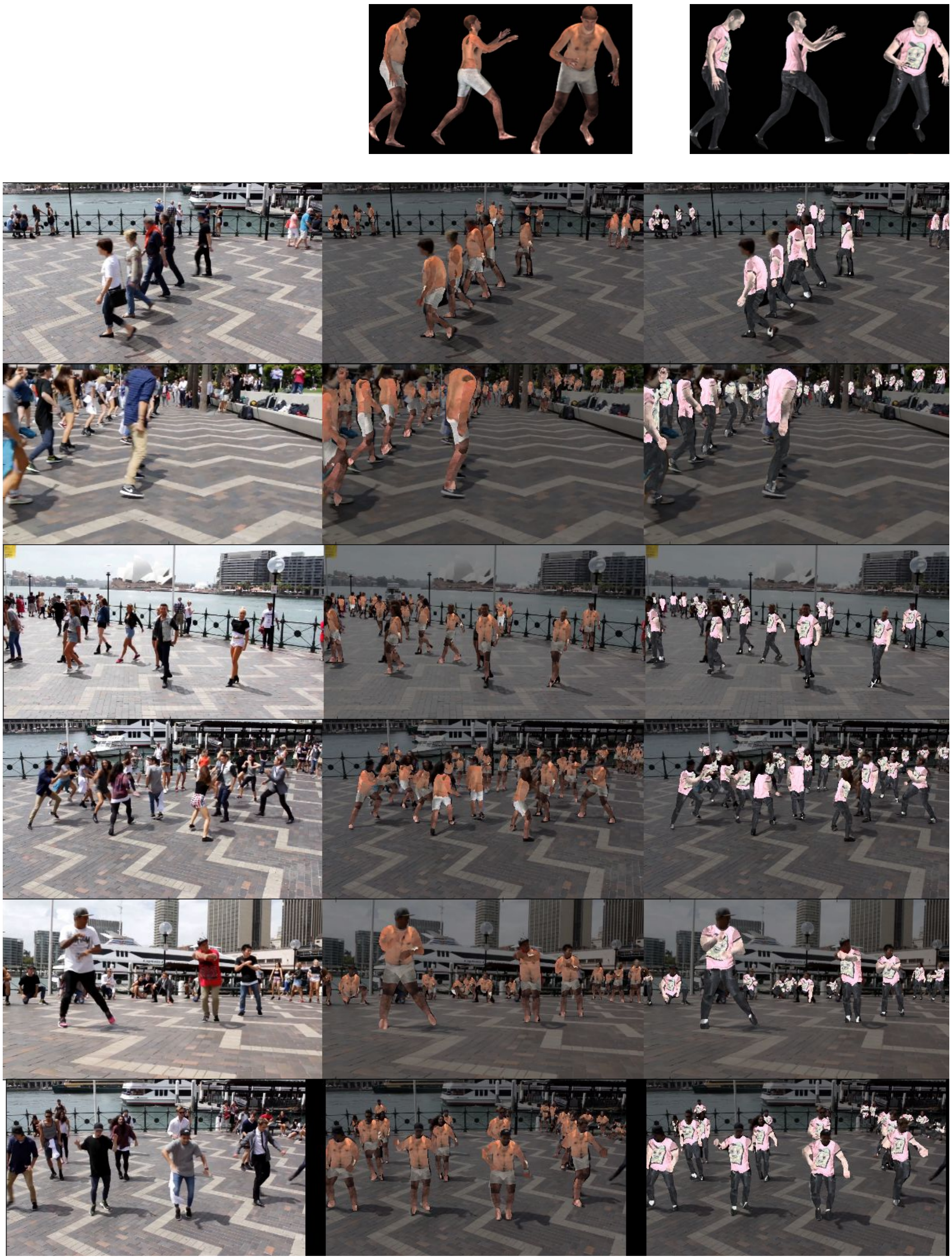


Figure 4.13: Qualitative results for texture transfer: The textures that are provided in the top row are mapped to image pixels based on estimated correspondences. The whole video can be seen at `http://densepose.org`.

Chapter 5

Dense Pose Transfer

In this work we integrate ideas from surface-based modeling with neural synthesis: we propose a combination of surface-based pose estimation and deep generative models that allows us to perform accurate pose transfer, i.e. synthesize a new image of a person based on a single image of that person and the image of a pose donor. We use a dense pose estimation system that maps pixels from both images to a common surface-based coordinate system, allowing the two images to be brought in correspondence with each other. We inpaint and refine the source image intensities in the surface coordinate system, prior to warping them onto the target pose. These predictions are fused with those of a convolutional predictive module through a neural synthesis module allowing for training the whole pipeline jointly end-to-end, optimizing a combination of adversarial and perceptual losses. We show that dense pose estimation is a substantially more powerful conditioning input than landmark-, or mask-based alternatives, and report systematic improvements over state of the art generators on DeepFashion and MVC datasets.

This work was done in collaboration with Natalia Neverova and is published at the European Conference on Computer Vision (ECCV 2018).

5.1 Introduction

Deep models have recently shown remarkable success in tasks such as face [Karras 2017], human [Lassner 2017a, Ma 2017, Siarohin 2018], or scene generation [Chen 2017b, Wang 2018b], collectively known as "neural synthesis". These results can look compellingly realistic, but their usefulness for graphics, or dataset augmentation tasks directly relates to the amount of control that one can exert on the generation process. Recent works have shown the possibility of manipulating image synthesis by controlling categorical attributes [Lample 2017, Lassner 2017a], low-dimensional parameters [Shu 2017], or layout constraints indicated by a conditioning input [Isola 2017, Chen 2017b, Wang 2018b, Lassner 2017a, Ma 2017, Siarohin 2018]. In this work we aspire to obtain a stronger hold of the image synthesis process by relying on surface-based object representations, similar to the ones used in graphics engines.

Our work is focused on the human body, where surface-based image understanding has been most recently unlocked [Loper 2015, Bogo 2016, Lassner 2017b, Varol 2017, Kanazawa 2018a], along with our contributions in this thesis, see Sec. 4. We build on the DensePose system described in Sec. 4, which allows us to interpret an image of a person in terms of a full-fledged surface model, namely perform

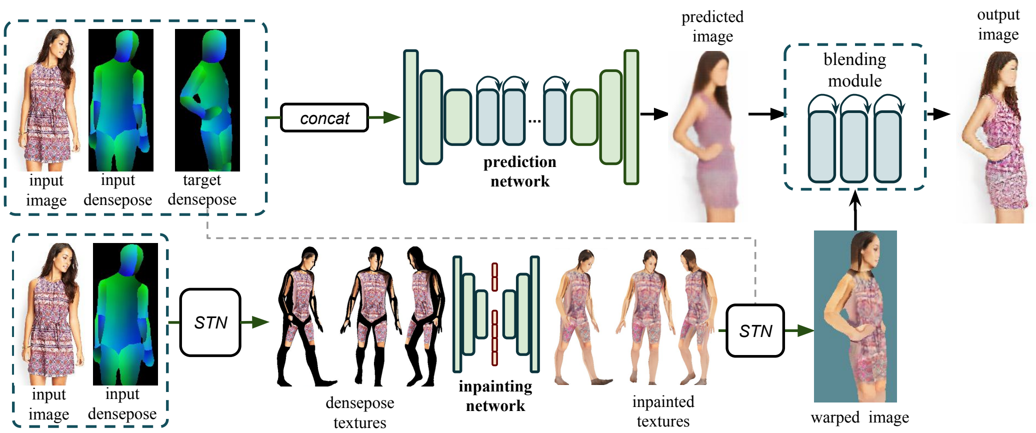


Figure 5.1: Overview of our two-stream pose transfer pipeline: given an input image and a target pose we use DensePose to drive the generation process. This is achieved through the complementary streams of (a) a data-driven predictive model, and (b) a surface-based model that warps the texture to UV-coordinates, interpolates on the surface, and warps back to the target image. A blending module exploits the complementary merits of these two streams to render the input image in the target pose.

"inverse graphics".

In this work we close the loop and perform image generation by rendering the same person in a new pose through surface-based neural synthesis. The target pose is indicated by an image of another person, and the DensePose system is used to associate the new photo with the common surface coordinates, and copy the appearance predicted there. We refer to this process as *surface coordination*, to indicate that it is accomplished by having both images parameterized in terms of a common, surface-based coordinate system.

The purely geometry-based synthesis process is on its own insufficient for realistic image generation: its performance can be compromised by inaccuracies of the DensePose system as well as by self-occlusions of the body surface in at least one of the two images. We account for occlusions by introducing an inpainting network that operates in the surface coordinate system, and combine its predictions with the outputs of a more traditional feedforward conditional synthesis module. These predictions are obtained independently, and compounded by a refinement module that is trained so as to optimize a combination of reconstruction, perceptual and adversarial losses.

We experiment on the DeepFashion dataset [Liu 2016c], and show that we can obtain results that are both qualitatively and quantitatively better than the latest state-of-the-art. Apart from the specific problem of pose transfer, the proposed combination of neural synthesis with surface-based representations is in our opinion also promising for the broader problems of virtual and augmented reality: the generation process is in a sense more transparent and easy to connect with the physical world, thanks to the underlying surface-based representation. In the more immediate future the task of pose transfer can be useful for dataset augmentation, as well as texture transfer applications like those showcased in Sec. 4, without however requiring the acquisition of a surface-level texture map.

5.2 Dense Pose Transfer

We develop our approach to pose transfer around the *DensePose* estimation system to associate every human pixel with its coordinates on a surface-based parameterization of the human body in an efficient, bottom-up manner. We exploit the DensePose outputs in two complementary ways, corresponding to the *warping model* and the *predictive model*, as shown in Fig. 5.1. The warping module uses DensePose-based surface correspondence and inpainting to generate a new view of the image, while the predictive module is a generic black-box generative model conditioned on the DensePose outputs for both the input and output images.

These modules have complementary merits: the *predictive model* successfully exploits the dense conditioning output to generate plausible images for familiar poses, delivering superior results to those obtained from sparse, landmark-based conditioning; at the same time, it cannot generalize to new poses, or transfer texture details. By contrast the *warping model* can preserve high quality details and textures, allows us to perform inpainting in a uniform, canonical coordinate system, and generalizes for free for a broad variety of body movements. However, its body-, rather than clothing-centered construction does not take into account hair, hanging clothes, and accessories. The best of both worlds is obtained by feeding the outputs of these two models into a *blending module* trained to combine and refine their predictions using a combination of reconstruction, adversarial, and perceptual losses.

Having outlined the overall architecture of our system, in Sec. 5.2.1 and Sec. 5.2.2 we present in some more detail our components, and then turn in Sec. 5.2.3 to the loss functions used in their training. A thorough description of architecture details is left to the supplemental material. We start by presenting the architecture of the predictive stream, and then turn to the surface-based stream, corresponding to the upper and lower rows of Fig. 5.1, respectively.

5.2.1 Predictive Stream

Dense pose estimation. The DensePose module is common to both streams and delivers dense correspondences between an image and a surface-based model of the human body. This system is trained discriminatively and provides a simple, feed-forward module for dense correspondence from an image to the human body surface. We omit further details, since we rely entirely on the DensePose system with minor differences in implementation described in Sec. 5.3.

Predictive model. This component is a conditional generative model that exploits the DensePose system results for pose transfer. Existing conditional models

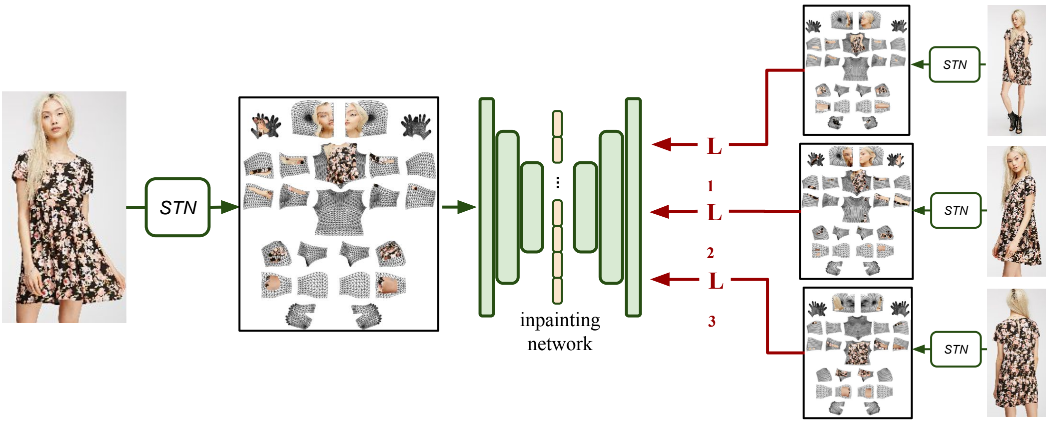


Figure 5.2: Supervision signals for pose transfer on the "surface coordination" stream: The input image on the left is warped to intrinsic surface coordinates through a spatial transformer network driven by DensePose. From this input, the Inpainting Autoencoder has to predict the appearance of the same person in different viewpoints, when also warped to intrinsic coordinates. The loss functions on the right penalize the reconstruction of the Autoencoder only on the observed parts of the texture map. This form of multi-view supervision acts like a surrogate for the (unavailable) appearance of the person on the full body surface. Similar supervision is used for the predictive stream, coming in the form of pairs of input and output poses of the same person with different poses.

indicate the target pose in the form of heat-maps from keypoint detectors [Ma 2017], or part segmentations [Lassner 2017a]. Here we condition on the concatenation of the input image and DensePose results for the input and target images, resulting in an input of dimension 256×256×9. This provides conditioning that is both global (part-classification), and point-level (continuous coordinates), allowing the remaining network to exploit a richer source of information.

The remaining architecture includes an encoder followed by a stack of residual blocks and an decoder at the end, along the lines of [Johnson 2016]. In more detail, this network comprises (a) a cascade of three convolutional layers that encode the 256×256×9 input into 64×64×256 activations, (b) a set of six residual blocks with 3×3×256×256 kernels, (c) a cascade of two deconvolutional and one convolutional layer that deliver an output of the same spatial resolution as the input. All intermediate convolutional layers have 3×3 filters and are followed by instance normalization [Ulyanov 2017] and ReLU activation. The last layer has tanh non-linearity and no normalization.

5.2.2 Surface coordination stream

Warping model. This module performs pose transfer by "coordinating" the input and the target image on the common surface UV-system.

The core of this component is a Spatial Transformer Network (STN) [Jaderberg 2015] that warps according to DensePose the image observations to the UV-coordinate system of each surface part; we use a grid with 256×256 UV points for each of the 24 surface parts, and perform scattered interpolation to handle the continuous values of the regressed UV coordinates. The inverse mapping from UV to the output image space is performed by a second STN with a bilinear kernel.

As shown in Fig. 5.3, a direct implementation of this module would often deliver poor results: the part of the surface that is visible on the source image is typically small, and can often be entirely non-overlapping with the part of the body that is visible on the target image. This is only exacerbated by DensePose failures or systematic errors around the part seams. These problems motivate the use of an inpainting network within the warping module, as detailed below.

Inpainting autoencoder. This model allows us to extrapolate the body appearance from the surface nodes populated by the STN to the remainder of the surface. Our setup requires a different approach to the one of other deep inpainting methods [Yeh 2017], because we never observe the full surface texture during training. We handle the partially-observed nature of our training signal by using a reconstruction loss that only penalizes the observed part of the UV map, and lets the network freely *guess* the remaining domain of the signal. In particular we use a masked ℓ_1 loss on the difference between the Autoencoder predictions and the target signals, where the masks indicate the visibility of the target signal.

We observed that by its own this does not urge the network to inpaint successfully; results substantially improve when we accompany every input with multiple supervision signals, as shown in Fig. 5.2, corresponding to UV-wrapped shots of the same person at different poses. This fills up a larger portion of the UV-space and forces the inpainting network to predict over the whole texture domain. As shown in Fig. 5.3, the inpainting process allows us to obtain a uniformly observed surface, which captures the appearance of skin and tight clothes, but does not account for hair, skirts, or apparel, since these are not accommodated by DensePose's surface model.

Blending module. As we have already mentioned, the two models described above have complementary merits. The blending module's objective is to combine their strengths and deliver a 'polished' result, as measured by the losses used for training. As such it no longer involves an encoder or decoder unit, but rather only contains two convolutional and three residual blocks that aim at combining the predictions and refining their results. The final refined prediction is obtained as a sum of the output of the predicted module and the residual term generated by the blending module.

5.2.3 Loss Functions

As shown in Fig. 5.1, the training set for our network comes in the form of pairs of input and target images, $\mathbf{x}$, $\mathbf{y}$ respectively, both of which are of the same person-clothing, but in different poses. Denoting by $\hat{\mathbf{y}} = G(\mathbf{x})$ the network's prediction, the difference between $\hat{\mathbf{y}}, \mathbf{y}$ can be measured through a multitude of loss terms, that

Figure 5.3: Warping module results: whole 3D model from a single image. For each sample, the top row shows interpolated textures obtained from DensePose predictions and projected on the surface of the 3D model of the body. The bottom row shows the same textures after inpainting in the UV space.

penalize different forms of deviation. We present them below for completeness, and refer to the original references for a more thorough analysis of their properties – we ablate their impact in practice in Sec. 5.3.

Reconstruction loss. To penalize reconstruction errors we use the common ℓ_1 distance between the two signals: $\|\hat{\mathbf{y}} - \mathbf{y}\|_1$. On its own it delivers blurry results, but is important for retaining the overall intensity levels. Apart from the outputs of the blending model, we use this loss also for the predictions of the warping and predictive modules of the networks, which amounts to performing deep supervision training.

Perceptual loss. As in Chen and Koltun [Chen 2017b], we use a VGG19 network pretrained for classification [Simonyan 2014b] as a feature extractor for both $\hat{\mathbf{y}}, \mathbf{y}$ and penalize the ℓ_2 distance of the respective intermediate feature activations Φ^v at 5 different network layers $v = 1, \ldots, N$:

$$\mathcal{L}_{\mathrm{p}}(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{v=1}^{N} \|\Phi^v(\mathbf{y}) - \Phi^v(\hat{\mathbf{y}})\|_2. \tag{5.1}$$

This loss penalizes differences in low- mid- and high-level feature statistics, captured by the respective network filters.

Style loss. As in [Johnson 2016], we use the Gram matrix criterion of [Gatys 2016] as an objective for training a feedforward network. Using the same notation as above, this criterion compute the Gram matrix of neuron activations delivered by the VGG network at layer v for an image $\mathbf{x}$: for input $\mathbf{x}$ at feature level v of network Φ are defined as follows:

$$\mathcal{G}^v(\mathbf{x})_{c,c'} = \sum_{h,w} \Phi_c^v(\mathbf{x})[h,w]\Phi_{c'}^v(\mathbf{x})[h,w] \tag{5.2}$$

where h and w are horizontal and vertical pixel coordinates and c and c' are feature maps of layer v. The style loss by the sum of the Frobenius norm of the difference between the per-layer Gram matrices $\mathcal{G}^v$ of the two inputs:

$$\mathcal{L}_{\text{style}}(\mathbf{y},\hat{\mathbf{y}}) = \sum_{v=1}^{B} \|\mathcal{G}^v(\mathbf{y}) - \mathcal{G}^v(\hat{\mathbf{y}})\|_F. \tag{5.3}$$

Adversarial loss. We use adversarial training to penalize any detectable differences between the generated and real samples. Since global structural properties are largely settled thanks to DensePose conditioning, we opt for the patch-GAN [Isola 2017] discriminator, which operates locally and picks up differences between texture patterns. As in [Isola 2017, Wang 2018b] we use a set of identical discriminators that process convolutionally the original image resolution and a downsampled version of it – the training loss is computed by evaluating each of the discriminators in a fully convolutional manner and summing the respective losses. The discriminator takes as an input $\mathbf{z}$, a combination of the source image and the DensePose results on the target image, and either the target image $\mathbf{y}$ (real) or the generated output (fake) $\hat{\mathbf{y}}$. We want fake samples to be indistinguishable from real ones – as such we optimize the following objective:

$$L_{\text{GAN}} = \underbrace{\frac{1}{2}\mathbb{E}_{\mathbf{z}}\left[l(D(\mathbf{z},\mathbf{y})-1)\right] + \frac{1}{2}\mathbb{E}_{\mathbf{z}}\left[l(D(\mathbf{z},\hat{\mathbf{y}}))\right]}_{\text{Discriminator}} + \underbrace{\frac{1}{2}\mathbb{E}_{\mathbf{z}}\left[l(D(G(\mathbf{z},\hat{\mathbf{y}})-1))\right]}_{\text{Generator}}, \tag{5.4}$$

where we use $l(x) = x^2$ as in the Least Squares GAN (LSGAN) work of [Mao 2017] for stability. To further stabilize the training, we adapt the discriminator feature matching strategy proposed in [Wang 2018b] and introduce an additional term analogous to Eq. (5.1) but defined on discriminator intermediate activations.

5.3 Experiments

Datasets

We perform our experiments on the DeepFashion dataset (In-shop Clothes Retrieval Benchmark) [Liu 2016c] that contains 52,712 images of fashion models demonstrating 13,029 clothing items in different poses. All images are provided at a resolution of 256×256 and contain people captured over a uniform background.

Following [Siarohin 2018] we select 12,029 clothes for training and the remaining 1,000 for testing. For the sake of direct comparison with state-of-the-art methods of keypoint-based image generation, we also remove all images where the keypoint detector of [Cao 2016] does not detect any body joints. This results in 140,110 training and 8,670 test pairs.

In the supplementary material we provide results on the large scale MVC dataset [Liu 2016a] that consists of 161,260 images of resolution 1920×2240 crawled from several online shopping websites and showing front, back, left, and right views for each clothing item.

Implementation details

DensePose estimator. We use a fully convolutional network, a ResNet-101 trained on cropped person instances from the COCO-DensePose dataset. To further improve the quality of predictions around the facial region, we also employ an additional ResNet-101 network based on DenseReg which is trained with strong 3DMM based [Booth 2016] supervision for learning dense correspondences for faces. The DenseReg system is trained on Menpo dataset [Zafeiriou 2017], which allows handling side-poses. We use the S^3FD face detector [Zhang 2017] to first detect the faces and get crops of normalized size on which DenseReg operates. The output of both body and face networks consists of 2D fields $\{I, U, V\}$ representing body segments (I) and U and V coordinates in coordinate spaces aligned with each of the semantic parts of the corresponding 3D model.
In our implementation, face textures of the source images are first warped and processed separately from the body and then combined at the input of the blending module - this step is omitted in figures for simplicity.

Training parameters. We train the network and its submodules with Adam optimizer with initial learning rate $2{\cdot}10^{-4}$ and β_1=0.5, β_2=0.999 (no weight decay). For speed, we pretrain the predictive module and the inpainting module separately and then train the blending network while finetuning the whole combined architecture end-to-end; DensePose network parameters remain fixed. In all experiments, the batch size is set to 8 and training proceeds for 40 epochs. The balancing weights λ between different losses in the blending step (described in Sec. 5.2.3) are set empirically to λ_{ℓ_1}=1, λ_{p}=0.5, λ_{style}=$5{\cdot}10^5$, λ_{GAN}=0.1.

Evaluation metrics

To the best of our knowledge there exists no criterion that would allow an adequate evaluation of the generated image quality from the perspective of both structural fidelity and photorealism. We therefore adopt a number of separate structural and perceptual metrics widely used in the community and report our joint performance on them.

Structure. The geometry of the generations is evaluated using the perception-correlated *Structural Similarity* metric (SSIM) [Wang 2004]. In this work we also exploit its multi-scale variant MS-SSIM [Wang 2003] to estimate the geometry of our predictions at a number of levels, from body structure to fine clothing textures.

Image realism. As in previous works we provide the values of *Inception scores*

Table 5.1: Quantitative comparison of model performance with the state-of-the-art methods on the DeepFashion dataset [Liu 2016c]. Our *best structure* model corresponds to the perceptual loss training, the *highest realism* model corresponds to the style loss training (more details are given in the text and Table 5.4). Our *balanced* model is trained using the full combination of losses.

Model	SSIM	IS	DS
Disentangled [Ma 2018]	0.614	3.29	–
VariGAN [Zhao 2018a]	0.620	3.03	–
G1+G2+D [Ma 2017]	0.762	3.09	–
DSC [Siarohin 2018]	0.761	3.39	0.966
Ours (best structure)	**0.796**	3.17	0.971
Ours (highest realism)	0.777	**3.67**	0.969
Ours (balanced)	0.785	3.61	0.971
Real data	*1.0*	*3.898*	*0.980*

(IS) [Salimans 2016]. However, as has repeatedly been noted in the literature, this metric is of limited relevance to the problem of within-class object generation, and we do not wish to draw strong conclusions from it. We have empirically observed instability and high variance of this metric with respect to the perceived quality of generations and structural similarity. We also note that the ground truth images from the DeepFashion dataset have an average IS of 3.9, which indicates low degree of 'realism' of this data according to the IS metric (for comparison, IS of CIFAR-10 is 11.2 [Salimans 2016] with best image generation methods achieving IS of 8.8 [Karras 2017]).

Finally, for the state-or-the-art comparison we perform additional evaluation using *detection scores (DS)* [Siarohin 2018] reflecting the similarity of the generated images to the *person* class. Detection scores correspond to the maximum of confidence of the PASCAL-trained SSD detector [Liu 2016b] in the person class taken over all bounding boxes detected in the image.

Comparison with the state-of-the-art

We compare performance of our framework with a number of recent methods proposed for the task of *keypoint guided* image generation or multi-view synthesis. Table 5.1 shows a significant advantage of our pipeline in terms of structural fidelity of obtained predictions. This holds for the whole range of tested network configurations and training setups (see Table 5.4). In terms of perceptional quality expressed through IS, the output generations of our models are of higher quality or en pair with the existing works. Some qualitative results of our method (corresponding to the *balanced* model in Table 5.1) and the best performing state-of-the-art approach [Siarohin 2018] are shown in Fig. 5.4.

Effectiveness of different body representations

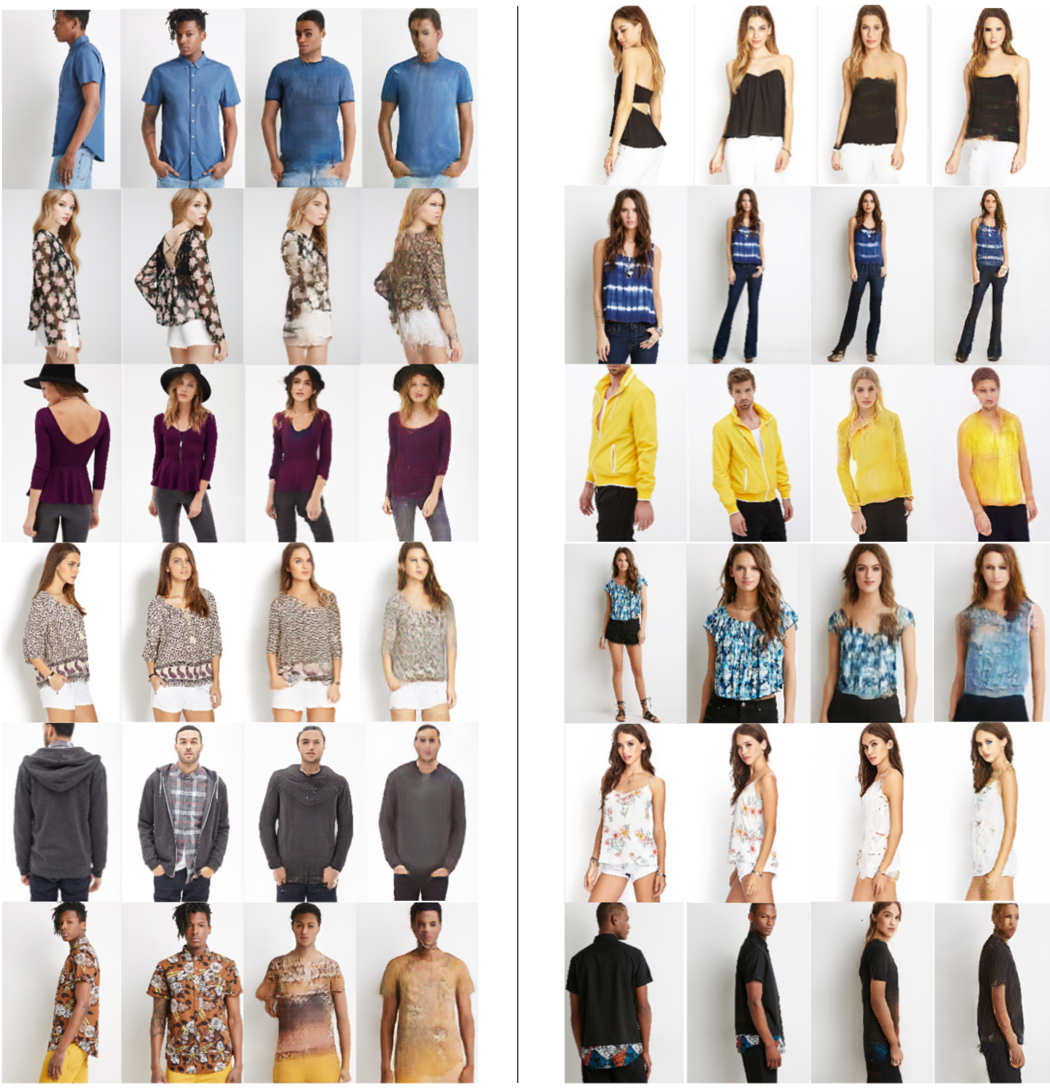


Figure 5.4: Qualitative comparison with the state-of-the-art Deformable GAN (DSC) method of [Siarohin 2018]. Each group shows the input, the target image, predictions by the DSC model [Siarohin 2018], predictions obtained with our full model. We observe that even though our cloth texture is occasionally not as sharp, we better retain face, gender, and even skin color information.

Figure 5.5: Typical failures of keypoint-based pose transfer frameworks (top) in comparison with DensePose conditioning (bottom) indicate disappearance of limbs, discontinuities, collapse of 3D geometry of the body into a single plane and confusion in ordering along the depth dimension.

Table 5.2: On effectiveness of different body representations as a ground for pose transfer. DensePose representation results in the highest structural quality of the predictions.

Model	SSIM	MS-SSIM	IS
Foreground mask	0.747	0.710	3.04
Body part segmentation	0.774	0.788	3.35
Body part segmentation, one-hot	0.776	0.791	3.22
Body keypoints, one-hot	0.762	0.774	3.09
DensePose $\{I, U, V\}$	**0.792**	**0.821**	3.09
DensePose $\{\text{one-hot}\, I, U, V\}$	0.782	0.799	3.32

We evaluate effectiveness of the DensePose representation as a ground for conditioning pose transfer frameworks compared to other more traditional body representations, such as background/foreground masks, body part segmentation maps or body landmarks.

As a segmentation map we take the index component of DensePose and evaluate two possible representations: either as a single plane with pixel values denoting their segment class, or one-hot encoding into a set of class specific binary masks. Accordingly, as a background/foreground mask, we simply take all pixels with positive DensePose segmentation indices. Finally, we follow [Siarohin 2018] and use the detector from [Cao 2016] to obtain body keypoints; along the lines of [Ma 2017, Siarohin 2018] we one-hot encode keypoints as gaussian heatmaps and provide them as inputs to the network.

In each case,we concatenate the source image with corresponding representation of the source and the target poses which results in 4 input planes for the mask, 4

Table 5.3: Contribution of each of the functional blocks of the framework

Model	SSIM	MS-SSIM	IS
predictive module only	0.792	0.821	3.09
predictive + blending (=self-refinement)	0.793	0.821	3.10
predictive + warping + blending	0.789	0.814	3.12
predictive + warping + inpainting + blending (full)	**0.796**	**0.823**	**3.17**

or 27 (one-hot) for segmentation maps and 21 for the keypoints. For simplicity we only train the predictive module, rather than the whole architecture.

The corresponding results shown in Table 5.2 demonstrate a clear advantage of fine-grained dense conditioning over the sparse, keypoint-based, or coarse, segmentation-based, representations. The one-hot encoding of the segmentation component does not significantly facilitate the training, possibly due to accompanying increase in the number of network parameters due to a higher dimensionality of the input.

Complementing these quantitative results, a number of typical failure cases of keypoint-based frameworks are demonstrated in Figure 5.5. We observe that these shortcomings are largely fixed by switching to the DensePose-based conditioning.

Ablation study on architectural choices

Table 5.3 shows contribution of each block (namely, predictive module, warping module, inpainting autoencoding) in the final model performance. For this experiments, we use only the reconstruction loss $\mathcal{L}_{\ell_1}$ (to avoid fluctuations in the performance due to instabilities of GAN training). As expected, including the warping branch in the generation pipeline results in better performance, which is further improved by including the inpainting in the UV space. Qualitatively, exploiting the inpainted representation has two advantages over the direct warping of the partially observed texture from the source pose to the target pose: first, it serves as an additional prior for the fusion pipeline, and, second, it also prevents the blending network from generating clearly visible sharp artifacts that otherwise appear on the boarders of partially observed segments of textures.

Ablation study on supervision objectives

Finally, we analyze the role of each of considered terms in the composite loss function used at the final stage of the training (see Table 5.4 for quantitative results and Fig. 5.6 for an illustration). Overall, the perceptual loss $\mathcal{L}_p$ turned out to be most correlated with the image structure and least correlated with the perceived realism, probably due to introduced textural artefacts. At the same time, the style loss $\mathcal{L}_{\text{style}}$ produces sharp and correctly textured patterns while hallucinating edges over uniform regions. Finally, adversarial training with the loss $\mathcal{L}_{\text{GAN}}$ tends to prioritize visual plausibility often disregarding information in the input. For these reasons, we combine all these complimentary supervision criteria with empirically chosen weights, as detailed in the training section.

Figure 5.6: Effects of training with different loss terms and their weighted combinations.

Table 5.4: Comparison of different loss terms used at the final stage of the training. Perceptual loss is best correlated with the structure, and style loss with IS. The combined model (last entry) provides an optimal balance between the extreme solutions.

Model	SSIM	MS-SSIM	IS
$\{\mathcal{L}_{\ell_1}, \mathcal{L}_p\}$	**0.791**	**0.822**	3.26
$\{\mathcal{L}_{\ell_1}, \mathcal{L}_{\text{style}}\}$	0.777	0.815	**3.67**
$\{\mathcal{L}_{\ell_1}, \mathcal{L}_p, \mathcal{L}_p\}$	0.784	0.820	3.41
$\{\mathcal{L}_{\ell_1}, \mathcal{L}_{\text{GAN}}\}$	0.771	0.807	3.39
$\{\mathcal{L}_{\ell_1}, \mathcal{L}_p, \mathcal{L}_{\text{GAN}}\}$	0.789	0.820	3.33
$\{\mathcal{L}_{\ell_1}, \mathcal{L}_{\text{style}}, \mathcal{L}_{\text{GAN}}\}$	0.787	0.820	3.32
$\{\mathcal{L}_{\ell_1}, \mathcal{L}_p, \mathcal{L}_{\text{style}}, \mathcal{L}_{\text{GAN}}\}$	**0.785**	**0.807**	**3.61**

5.4 Conclusion

In this work we have introduced a two-stream architecture for pose transfer that exploits the power of dense human pose estimation. We have shown that dense pose estimation is a clearly superior conditioning signal for data-driven human pose estimation, and also facilitates the formulation of the pose transfer problem in its natural, body-surface parameterization through inpainting. In future work we intend to further pursue the potential of this method for photorealistic image synthesis [Karras 2017, Chen 2017b].

Chapter 6

Conclusion and Future Work

Within the thesis, we have pushed further the envelope of tasks that can be addressed by CNNs and considered a task that lies at the end of the 'location detail' spectrum. We have introduced a regression-based approach to establishing dense correspondences between image pixels and object templates. We have described a customized pipeline to collect ground truth image-to-surface annotations for the human body, allowing inference of dense correspondences from RGB images for the first time. Through live demonstrations, we have presented that our dense correspondence systems can perform considerably well in real time using a single GPU.

We have shown that the image-to-template correspondences proposed in this thesis can be used to solve a host of problems, such as texture transfer, by using the template as a proxy. Through our technical contributions, we have reported state-of-the-art results in a multitude of computer vision tasks: facial landmark localization, facial part segmentation, 3D human joint localization, monocular dense correspondence estimation and human image synthesis.

Despite these advances, we are still far from recovering the entirety of the information one can elicit from an image. A limitation of the proposed dense human pose estimation system is that it establishes correspondences to the human body and ignores the clothes. This limits some of the potential use-cases for loosely clothed humans in images, e.g. skirts, dresses. It is also important to note that DensePose does not output the 3D reconstruction of the shape, particularly, correspondence for invisible parts of the body and the depth is unknown.

The research presented in this thesis is only a stepping stone to such a full-blown image understanding. We describe below directions for further research, stemming from the contributions of the thesis.

6.1 3D Human Body Shape Reconstruction In-the-wild

The conventional way [Vetter 1997b, Blanz 2003b] to fit morphable models is to optimize model parameters such that the model is in alignment with the object in the image, see Sec. 1.2.2. There are recent works that aim at fitting the statistical deformable model of the human body to images, an example is [Kanazawa 2018b], for a more detailed review please see Sec. 1.2.2.2. These share the common goal of predicting model parameters such that the model joints are in alignment with ground truth joints. The objective function for fitting can be enriched with dense correspondences.

As manual annotations or bottom-up predictions, dense correspondences well complement existing cues such as 2D and 3D joints. A prominent future research direction involves the incorporation of dense correspondences into the objective function of model-based 3D human body shape reconstruction systems for better surface alignment.

6.2 Human Image Synthesis

Photorealism of synthesized images by generative adversarial networks are getting increasingly better as the training strategies are enhanced [Karras 2017,Brock 2018]. Synthesizing humans in different poses or with modified appearances is an interesting problem with applications in augmented reality or fashion. We show in Chap. 5 that dense correspondences can be utilized to improve performance for such systems over baselines as [Lassner 2017a, Ma 2017, Siarohin 2018]. Recently, [Wang 2018a] has shown that it is possible to synthesize new videos of a person given target dense correspondences, using a system trained from videos of that specific person in fixed clothing.

The current state-of-the-art systems are far from generating images with a desirable level of photorealism when it comes to generalizing to multiple people and clothes, eg. Deepfashion dataset [Liu 2016c]. DensePose or even a perfectly fitted 3D morphable model provides correspondences to the human body and not the clothes. Investigating representations for clothed regions in relation with the human body geometry is an important future direction.

6.3 Extension to More Objects

Within the thesis we have demonstrated dense correspondence results on the human body, face and ear. A clear direction forward is extending the repertoire of the proposed dense correspondence systems. The straightforward extension is to design a template space for another deformable object. One example would be four-legged mammals, for which there exists a statistical deformable model [Zuffi 2017, Zuffi 2018]. Moreover, the use cases of the proposed framework can be extended to many categories, including man-made object categories.

Our framework can address the setting where the variation between different samples of the same object category is modeled as deformations from a template. An inherent challenge is the topological inconsistency between samples from the same object category. For instance, some cars have four doors whereas some have two, or some cars have spoilers and some do not. It is not straightforward to have a single canonical template to represent correspondences between any two cars. Similar challenges occur in establishing dense correspondences among 3D rigid shapes, eg. [Kim 2012,Huang 2018] or co-segmentation, eg. [Huang 2011,Sidi 2011], as detailed in surveys of [Mitra 2013,Xu 2017]. There are recent efforts to get region annotations in large 3D shape collections [Yi 2016]. There are also efforts to collect ground truth that respects the hierarchical nature of semantics [Yi 2017, Mo 2018]. These allow not only a geometric but also a semantic and functional partitioning of shapes. The local geometry on such parts can be represented using deformation-free coordinate systems.

A future research direction is the investigation of data collection pipelines and systems that establish dense correspondences between RGB images and coordinates defined on hierarchical part templates for many objects, including man-made ones.

6.4 Unsupervised / Weakly Supervised Learning

Despite the well-optimized annotation pipeline, the cost of the DensePose-COCO dataset is approximately 30,000$. The human body is a particularly important category with many crucial applications and merits the special treatment of hand engineering an annotation system and collecting expensive annotations. However, it is not feasible to repeat these steps for hundreds of common object classes. This motivates establishing correspondences in a weakly-supervised or even unsupervised manner. Recently, [Thewlis 2017] shows that one can align sets of images on a fixed coordinate system using the equivariance principle with no supervision. [Shu 2018] shows that one can learn to generate deformation fields and deformation-free appearance images using Deforming Auto-Encoders. These approaches work well for the human face, which has a quite simple geometry with no articulations. Recently, [Kanazawa 2018c] shows that using segmentations, landmarks and symmetry assumption one can form a 3D morphable model of an object. This was done using a neural renderer module, [Kato 2018], that allows differentiable image formation from a mesh and a texture.

Weakly supervised and unsupervised dense correspondence estimation could be instrumental in scaling the number of objects. Another potential direction of research that falls under this category is the discovery of 3D shape for the human face, human body and hand from weak supervision signals such as 2D joints and motion.

6.5 Action Recognition

One of the most significant fields of research in computer vision is action recognition, which deals with classifying an action in a given video. Recent works make use of motion-encoding inputs to their systems, most commonly known with the two-stream convolutions with an optical flow input branch [Simonyan 2014a]. Some recent alternatives are the difference of consecutive frames [Wang 2016] or images representative of motion dynamics [Bilen 2016]. Since many actions are tied to the motion of humans in the scene, it is intuitive to incorporate human pose information. The human pose was shown to help action recognition while extracting features, e.g. [Chéron 2015, Zolfaghari 2017, Choutas 2018] or within a multi-task setting, e.g. [Luvizon 2018].

A potential research direction is to investigate if incorporating the proposed dense human pose estimation in action recognition pipelines would lead to an even further improvement compared to the sparse landmark based human pose. Since DensePose is defined on the image domain, it is straightforward to add it as an additional input stream. It would also be possible to adopt 'DensePose-flow' that encodes human body motion based on DensePose as an alternative to generic optical flow that typically enforces brightness consistency.

Bibliography

[Agarwal 2006a] Ankur Agarwal and Bill Triggs. *A local basis representation for estimating human pose from cluttered images.* In Asian Conference on Computer Vision, pages 50–59. Springer, 2006. (Cited on page 8.)

[Agarwal 2006b] Ankur Agarwal and Bill Triggs. *Recovering 3D human pose from monocular images.* IEEE transactions on pattern analysis and machine intelligence, vol. 28, no. 1, pages 44–58, 2006. (Cited on page 13.)

[Akhter 2015] Ijaz Akhter and Michael J Black. *Pose-conditioned joint angle limits for 3D human pose reconstruction.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1446–1455, 2015. (Cited on page 13.)

[Allen 2002] Brett Allen, Brian Curless and Zoran Popović. *Articulated body deformation from range scan data.* In ACM Transactions on Graphics (TOG), volume 21, pages 612–619. ACM, 2002. (Cited on page 19.)

[Allen 2003] Brett Allen, Brian Curless and Zoran Popović. *The space of human body shapes: reconstruction and parameterization from range scans.* In ACM transactions on graphics (TOG), volume 22, pages 587–594. ACM, 2003. (Cited on page 19.)

[Allen 2006] Brett Allen, Brian Curless, Zoran Popović and Aaron Hertzmann. *Learning a correlated model of identity and pose-dependent body shape variation for real-time synthesis.* In Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation, pages 147–156. Eurographics Association, 2006. (Cited on page 20.)

[Amberg 2007] Brian Amberg, Sami Romdhani and Thomas Vetter. *Optimal step nonrigid ICP algorithms for surface registration.* In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, pages 1–8. IEEE, 2007. (Cited on page 17.)

[Amit 1991] Yali Amit, Ulf Grenander and Mauro Piccioni. *Structural image restoration through deformable templates.* Journal of the American Statistical Association, vol. 86, no. 414, pages 376–387, 1991. (Cited on pages 15 and 25.)

[Andriluka 2009] Mykhaylo Andriluka, Stefan Roth and Bernt Schiele. *Pictorial structures revisited: People detection and articulated pose estimation.* In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 1014–1021. IEEE, 2009. (Cited on page 12.)

[Andriluka 2014] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler and Bernt Schiele. *2D Human Pose Estimation: New Benchmark and State of the Art Analysis*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2014. (Cited on page 14.)

[Anguelov 2005] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers and James Davis. *SCAPE: shape completion and animation of people*. In ACM transactions on graphics (TOG), volume 24, pages 408–416. ACM, 2005. (Cited on pages 18 and 19.)

[Antonakos 2015] Epameinondas Antonakos, Joan Alabort-i-Medina, Georgios Tzimiropoulos and Stefanos Zafeiriou. *Feature-Based Lucas-Kanade and Active Appearance Models*. IEEE Transactions on Image Processing, vol. 24, no. 9, September 2015. (Cited on page 34.)

[Bai 2017] Min Bai and Raquel Urtasun. *Deep watershed transform for instance segmentation*. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2858–2866. IEEE, 2017. (Cited on page 11.)

[Barrón 2001] Carlos Barrón and Ioannis A Kakadiaris. *Estimating anthropometry and pose from a single uncalibrated image*. Computer Vision and Image Understanding, vol. 81, no. 3, pages 269–284, 2001. (Cited on page 13.)

[Bas 2017] Anil Bas, Patrik Huber, William AP Smith, Muhammad Awais and Josef Kittler. *3d morphable models as spatial transformer networks*. In Proc. ICCV Workshop on Geometry Meets Deep Learning, pages 904–912, 2017. (Cited on page 18.)

[Belagiannis 2014] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab and Slobodan Ilic. *3D pictorial structures for multiple human pose estimation*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1669–1676, 2014. (Cited on pages 14 and 18.)

[Belagiannis 2017] Vasileios Belagiannis and Andrew Zisserman. *Recurrent human pose estimation*. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pages 468–475. IEEE, 2017. (Cited on page 12.)

[Bilen 2016] Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi and Stephen Gould. *Dynamic image networks for action recognition*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3034–3042, 2016. (Cited on page 85.)

[Blanz 1999] Volker Blanz and Thomas Vetter. *A morphable model for the synthesis of 3D faces*. In Proceedings of the 26th annual conference on Computer graphics and interactive techniques, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999. (Cited on pages 4, 16, 17 and 18.)

[Blanz 2003a] Volker Blanz and Thomas Vetter. *Face recognition based on fitting a 3D morphable model*. IEEE Transactions on pattern analysis and machine intelligence, vol. 25, no. 9, pages 1063–1074, 2003. (Cited on page 16.)

[Blanz 2003b] Volker Blanz and Thomas Vetter. *Face Recognition Based on Fitting a 3D Morphable Model*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 9, pages 1063–1074, 2003. (Cited on page 83.)

[Bogo 2016] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero and Michael J. Black. *Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image*. In Computer Vision – ECCV 2016, Lecture Notes in Computer Science. Springer International Publishing, October 2016. (Cited on pages 20, 39, 61, 62, 63 and 69.)

[Booth 2014] James Booth and Stefanos Zafeiriou. *Optimal UV spaces for facial morphable model construction*. In 2014 IEEE International Conference on Image Processing. IEEE, 2014. (Cited on page 26.)

[Booth 2016] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah and David Dunaway. *A 3D Morphable Model learnt from 10,000 faces*. In Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition. IEEE, 2016. (Cited on pages 17, 26 and 76.)

[Boussaid 2014] Haithem Boussaid and Iasonas Kokkinos. *Fast and exact: ADMM-based discriminative shape segmentation with loopy part models*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4058–4065, 2014. (Cited on pages 42, 43, 44 and 45.)

[Boyd 2011] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato and Jonathan Eckstein. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Foundations and Trends® in Machine Learning, vol. 3, no. 1, pages 1–122, 2011. (Cited on pages 42 and 43.)

[Brand 1999] Matthew Brand. *Shadow puppetry*. In Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, volume 2, pages 1237–1244. IEEE, 1999. (Cited on page 13.)

[Bristow 2015] Hilton Bristow, Jack Valmadre and Simon Lucey. *Dense semantic correspondence where every pixel is a classifier*. In Proceedings of the IEEE International Conference on Computer Vision, pages 4024–4031, 2015. (Cited on page 16.)

[Brock 2018] Andrew Brock, Jeff Donahue and Karen Simonyan. *Large scale gan training for high fidelity natural image synthesis*. arXiv preprint arXiv:1809.11096, 2018. (Cited on page 84.)

[Burl 1998] Michael C Burl, Markus Weber and Pietro Perona. *A probabilistic approach to object recognition using local photometry and global geometry*.

In European conference on computer vision, pages 628–641. Springer, 1998. (Cited on page 15.)

[Cao 2014] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong and Kun Zhou. *Facewarehouse: A 3d facial expression database for visual computing.* IEEE Transactions on Visualization and Computer Graphics, vol. 20, no. 3, 2014. (Cited on pages 17 and 30.)

[Cao 2016] Zhe Cao, Tomas Simon, Shih-En Wei and Yaser Sheikh. *Realtime multi-person 2d pose estimation using part affinity fields.* arXiv preprint arXiv:1611.08050, 2016. (Cited on pages 12, 76 and 79.)

[Carreira 2016] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki and Jitendra Malik. *Human pose estimation with iterative error feedback.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4733–4742, 2016. (Cited on page 12.)

[Cashman 2013] Thomas J Cashman and Andrew W Fitzgibbon. *What shape are dolphins? building 3d morphable models from 2d images.* IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 1, pages 232–244, 2013. (Cited on page 16.)

[Čech 2016] Jan Čech, Vojtěch Franc, Michal Uřičář and Jiří Matas. *Multi-view facial landmark detection by using a 3D shape model.* Image and Vision Computing, vol. 47, 2016. (Cited on page 34.)

[Chen 2013] Yinpeng Chen, Zicheng Liu and Zhengyou Zhang. *Tensor-based human body modeling.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 105–112, 2013. (Cited on page 20.)

[Chen 2014] Xianjie Chen and Alan L Yuille. *Articulated pose estimation by a graphical model with image dependent pairwise relations.* In Advances in neural information processing systems, pages 1736–1744, 2014. (Cited on pages 12 and 42.)

[Chen 2016a] Ching-Hang Chen and Deva Ramanan. *3D Human Pose Estimation= 2D Pose Estimation+ Matching.* arXiv preprint arXiv:1612.06524, 2016. (Cited on pages 14 and 39.)

[Chen 2016b] Dong Chen, Gang Hua, Fang Wen and Jian Sun. *Supervised Transformer Network for Efficient Face Detection.* In European Conference on Computer Vision, 2016. (Cited on page 4.)

[Chen 2016c] Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or and Baoquan Chen. *Synthesizing training images for boosting human 3d pose estimation.* In 3D Vision (3DV), 2016 Fourth International Conference on, pages 479–488. IEEE, 2016. (Cited on pages 14 and 61.)

[Chen 2017a] Ching-Hang Chen and Deva Ramanan. *3d human pose estimation= 2d pose estimation+ matching*. In CVPR, volume 2, page 6, 2017. (Cited on page 14.)

[Chen 2017b] Qifeng Chen and Vladlen Koltun. *Photographic Image Synthesis with Cascaded Refinement Networks*. In ICCV, 2017. (Cited on pages 69, 74 and 81.)

[Chen 2017c] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu and Jian Sun. *Cascaded pyramid network for multi-person pose estimation*. arXiv preprint arXiv:1711.07319, 2017. (Cited on page 12.)

[Chen 2018a] Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang and Hartwig Adam. *Masklab: Instance segmentation by refining object detection with semantic and direction features*. In CVPR, volume 2, 2018. (Cited on page 10.)

[Chen 2018b] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy and Alan L Yuille. *Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs*. IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 4, pages 834–848, 2018. (Cited on pages 10, 27, 30, 31 and 52.)

[Chéron 2015] Guilhem Chéron, Ivan Laptev and Cordelia Schmid. *P-cnn: Pose-based cnn features for action recognition*. In Proceedings of the IEEE international conference on computer vision, pages 3218–3226, 2015. (Cited on page 85.)

[Choutas 2018] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud and Cordelia Schmid. *PoTion: Pose MoTion Representation for Action Recognition*. In CVPR 2018, 2018. (Cited on page 85.)

[Chrysos 2015] G. Chrysos, E. Antonakos, S. Zafeiriou and P. Snape. *Offline Deformable Face Tracking in Arbitrary Videos*. In Proceedings of IEEE International Conference on Computer Vision, 300 Videos in the Wild (300-VW): Facial Landmark Tracking in-the-Wild Challenge & Workshop (ICCVW'15), Santiago, Chile, December 2015. (Cited on page 35.)

[Cootes 1992] Timothy F Cootes, Christopher J Taylor, David H Cooper and Jim Graham. *Training models of shape from sets of examples*. In BMVC92, pages 9–18. Springer, 1992. (Cited on pages 16 and 17.)

[Cootes 2001] Timothy F Cootes, Gareth J Edwards and Christopher J Taylor. *Active appearance models*. IEEE Transactions on Pattern Analysis & Machine Intelligence, no. 6, pages 681–685, 2001. (Cited on pages 4, 16, 17 and 25.)

[Dai 2015] Jifeng Dai, Kaiming He and Jian Sun. *Convolutional feature masking for joint object and stuff segmentation*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3992–4000, 2015. (Cited on page 10.)

[Dai 2016a] Jifeng Dai, Kaiming He and Jian Sun. *Instance-aware semantic segmentation via multi-task network cascades.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3150–3158, 2016. (Cited on page 10.)

[Dai 2016b] Jifeng Dai, Yi Li, Kaiming He and Jian Sun. *R-fcn: Object detection via region-based fully convolutional networks.* In Advances in neural information processing systems, pages 379–387, 2016. (Cited on page 9.)

[Dalal 2005] Navneet Dalal and Bill Triggs. *Histograms of oriented gradients for human detection.* In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 1, pages 886–893. IEEE, 2005. (Cited on pages 8 and 9.)

[Dantone 2013] Matthias Dantone, Juergen Gall, Christian Leistner and Luc Van Gool. *Human pose estimation using body parts dependent joint regressors.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3041–3048, 2013. (Cited on page 12.)

[De Brabandere 2017] Bert De Brabandere, Davy Neven and Luc Van Gool. *Semantic instance segmentation with a discriminative loss function.* arXiv preprint arXiv:1708.02551, 2017. (Cited on page 11.)

[Deng 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei. *Imagenet: A large-scale hierarchical image database.* In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 248–255. Ieee, 2009. (Cited on page 53.)

[Deng 2016] Jiankang Deng, Qingshan Liu, Jing Yang and Dacheng Tao. *M 3 CSR: Multi-view, multi-scale and multi-component cascade shape regression.* Image and Vision Computing, vol. 47, 2016. (Cited on page 34.)

[Duetscher 2000] Jonathan Duetscher, Andrew Blake and Ian Reid. *Articulated body motion capture by annealed particle filtering.* In cvpr, page 2126. IEEE, 2000. (Cited on page 18.)

[Durer 1534] Albrecht Durer. Four books on human proportion(vier bucher von menschlicher proportion). Hieronymus Andreae, called Formschneyder, 1534. (Cited on page 15.)

[Edwards 1998] Gareth J Edwards, Timothy F Cootes and Christopher J Taylor. *Face recognition using active appearance models.* In European conference on computer vision, pages 581–595. Springer, 1998. (Cited on page 16.)

[Eichner 2009] Marcin Eichner, Vittorio Ferrari and S Zurich. *Better Appearance Models for Pictorial Structures.* In Bmvc, volume 2, page 5, 2009. (Cited on page 12.)

[Everingham 2015] Mark Everingham, S. M. Ali Eslami, Luc J. Van Gool, Christopher K. I. Williams, John M. Winn and Andrew Zisserman. *The Pascal Visual Object Classes Challenge: A Retrospective.* International Journal of Computer Vision, vol. 111, no. 1, pages 98–136, 2015. (Cited on page 53.)

[Fan 2016] Haoqiang Fan and Erjin Zhou. *Approaching human level facial landmark localization by deep learning.* Image and Vision Computing, vol. 47, 2016. (Cited on page 34.)

[Fang 2017] Haoshu Fang, Shuqin Xie, Yu-Wing Tai and Cewu Lu. *Rmpe: Regional multi-person pose estimation.* In The IEEE International Conference on Computer Vision (ICCV), volume 2, 2017. (Cited on page 12.)

[Fathi 2017] Alireza Fathi, Zbigniew Wojna, Vivek Rathod, Peng Wang, Hyun Oh Song, Sergio Guadarrama and Kevin P Murphy. *Semantic instance segmentation via deep metric learning.* arXiv preprint arXiv:1703.10277, 2017. (Cited on page 11.)

[Felzenszwalb 2005] Pedro F Felzenszwalb and Daniel P Huttenlocher. *Pictorial structures for object recognition.* International journal of computer vision, vol. 61, no. 1, pages 55–79, 2005. (Cited on pages 12, 15, 18 and 42.)

[Felzenszwalb 2008] Pedro Felzenszwalb, David McAllester and Deva Ramanan. *A discriminatively trained, multiscale, deformable part model.* In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pages 1–8. IEEE, 2008. (Cited on pages 9 and 18.)

[Felzenszwalb 2010] Pedro F Felzenszwalb, Ross B Girshick, David McAllester and Deva Ramanan. *Object detection with discriminatively trained part-based models.* IEEE transactions on pattern analysis and machine intelligence, vol. 32, no. 9, pages 1627–1645, 2010. (Not cited.)

[Fischler 1973] Martin A Fischler and Robert A Elschlager. *The representation and matching of pictorial structures.* IEEE Transactions on computers, vol. 100, no. 1, pages 67–92, 1973. (Cited on pages 9 and 15.)

[Frey 2003] Brendan J. Frey and Nebojsa Jojic. *Transformation-invariant clustering using the EM algorithm.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 1, pages 1–17, 2003. (Cited on page 16.)

[Ganin 2015] Yaroslav Ganin and Victor S. Lempitsky. *Unsupervised Domain Adaptation by Backpropagation.* In ICML, 2015. (Cited on page 62.)

[Gatys 2016] L. A. Gatys, A. S. Ecker and M. Bethge. *A neural algorithm of artistic style.* In CVPR, 2016. (Cited on page 75.)

[Gavrila 1996] Dariu M Gavrila and Larry S Davis. *3-D model-based tracking of humans in action: a multi-view approach.* In Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 73–80. IEEE, 1996. (Cited on pages 18 and 19.)

[Ghezelghieh 2016] Mona Fathollahi Ghezelghieh, Rangachar Kasturi and Sudeep Sarkar. *Learning camera viewpoint using CNN to improve 3D body pose estimation.* In 3DV, 2016. (Cited on page 61.)

[Girshick 2014] Ross Girshick, Jeff Donahue, Trevor Darrell and Jitendra Malik. *Rich feature hierarchies for accurate object detection and semantic segmentation.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 580–587, 2014. (Cited on page 9.)

[Girshick 2015] Ross Girshick. *Fast r-cnn.* In Proceedings of the IEEE international conference on computer vision, pages 1440–1448, 2015. (Cited on pages 9, 14, 28 and 41.)

[Grauman 2003] Kristen Grauman, Gregory Shakhnarovich and Trevor Darrell. *Inferring 3d structure with a statistical image-based shape model.* In null, page 641. IEEE, 2003. (Cited on page 13.)

[Grenander 1976] Ulf Grenander. *Pattern Synthesis: Lectures in Pattern Theory, vol. 1.* Applied Mathematical Sciences, vol. 18, 1976. (Cited on page 15.)

[Ham 2016] Bumsub Ham, Minsu Cho, Cordelia Schmid and Jean Ponce. *Proposal flow.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3475–3484, 2016. (Cited on page 16.)

[Han 2017] Kai Han, Rafael S Rezende, Bumsub Ham, Kwan-Yee K Wong, Minsu Cho, Cordelia Schmid and Jean Ponce. *SCNet: Learning semantic correspondence.* arXiv preprint arXiv:1705.04043, 2017. (Cited on page 16.)

[Hariharan 2014] Bharath Hariharan, Pablo Arbeláez, Ross Girshick and Jitendra Malik. *Simultaneous detection and segmentation.* In European Conference on Computer Vision, pages 297–312. Springer, 2014. (Cited on page 10.)

[Hariharan 2015] Bharath Hariharan, Pablo Arbeláez, Ross Girshick and Jitendra Malik. *Hypercolumns for object segmentation and fine-grained localization.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 447–456, 2015. (Cited on page 10.)

[Harzallah 2009] Hedi Harzallah, Frédéric Jurie and Cordelia Schmid. *Combining efficient object localization and image classification.* In Computer Vision, 2009 IEEE 12th International Conference on, pages 237–244. IEEE, 2009. (Cited on page 9.)

[Hasler 2009] Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo Rosenhahn and H-P Seidel. *A statistical model of human pose and body shape.* In Computer Graphics Forum, volume 28, pages 337–346. Wiley Online Library, 2009. (Cited on page 20.)

[Hasler 2010] Nils Hasler, Thorsten Thormählen, Bodo Rosenhahn and Hans-Peter Seidel. *Learning skeletons for shape and pose.* In Proceedings of the 2010 ACM SIGGRAPH symposium on Interactive 3D Graphics and Games, pages 23–30. ACM, 2010. (Cited on page 20.)

[He 2004] Xuming He, Richard S Zemel and Miguel Á Carreira-Perpiñán. *Multiscale conditional random fields for image labeling.* In Computer vision and pattern recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE computer society conference on, volume 2, pages II–II. IEEE, 2004. (Cited on page 10.)

[He 2014] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. *Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition.* In ECCV, 2014. (Cited on page 58.)

[He 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. *Identity mappings in deep residual networks.* In European conference on computer vision, pages 630–645. Springer, 2016. (Cited on page 30.)

[He 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár and Ross Girshick. *Mask r-cnn.* In Computer Vision (ICCV), 2017 IEEE International Conference on, pages 2980–2988. IEEE, 2017. (Cited on pages 10, 12, 51, 52, 57, 58, 59 and 60.)

[Hesse 2018] Nikolas Hesse, Sergi Pujades, Javier Romero, Michael J Black, Christoph Bodensteiner, Michael Arens, Ulrich G Hofmann, Uta Tacke, Mijna Hadders-Algra, Raphael Weinberger*et al.* *Learning an Infant Body Model from RGB-D Data for Accurate Full Body Motion Analysis.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 792–800. Springer, 2018. (Cited on page 20.)

[Hirshberg 2012] David A Hirshberg, Matthew Loper, Eric Rachlin and Michael J Black. *Coregistration: Simultaneous alignment and modeling of articulated 3D shape.* In European conference on computer vision, pages 242–255. Springer, 2012. (Cited on page 20.)

[Hogg 1983] David Hogg. *Model-based vision: a program to see a walking person.* Image and Vision computing, vol. 1, no. 1, pages 5–20, 1983. (Cited on page 18.)

[Holschneider 1990] Matthias Holschneider, Richard Kronland-Martinet, Jean Morlet and Ph Tchamitchian. *A real-time algorithm for signal analysis with the help of the wavelet transform.* In Wavelets, pages 286–297. Springer, 1990. (Cited on page 10.)

[Huang 2011] Qixing Huang, Vladlen Koltun and Leonidas Guibas. *Joint shape segmentation with linear programming.* In ACM transactions on graphics (TOG), volume 30, page 125. ACM, 2011. (Cited on page 84.)

[Huang 2018] Haibin Huang, Evangelos Kalogerakis, Siddhartha Chaudhuri, Duygu Ceylan, Vladimir G Kim and Ersin Yumer. *Learning Local Shape Descriptors from Part Correspondences with Multiview Convolutional Networks.* ACM Transactions on Graphics (TOG), vol. 37, page 6, 2018. (Cited on page 84.)

[Huber 2016] Patrik Huber, Guosheng Hu, Rafael Tena, Pouria Mortazavian, P Koppen, William J Christmas, Matthias Ratsch and Josef Kittler. *A multiresolution 3d morphable face model and fitting framework.* In Proceedings of the 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, 2016. (Cited on page 18.)

[Insafutdinov 2016] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka and Bernt Schiele. *Deepercut: A deeper, stronger, and faster multi-person pose estimation model.* In European Conference on Computer Vision, pages 34–50. Springer, 2016. (Cited on pages 12, 45 and 63.)

[Ionescu 2014a] Catalin Ionescu, Joao Carreira and Cristian Sminchisescu. *Iterated second-order label sensitive pooling for 3d human pose estimation.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1661–1668, 2014. (Cited on page 14.)

[Ionescu 2014b] Catalin Ionescu, Dragos Papava, Vlad Olaru and Cristian Sminchisescu. *Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 7, pages 1325–1339, jul 2014. (Cited on pages 4, 13, 14 and 46.)

[Iqbal 2016] Umar Iqbal and Juergen Gall. *Multi-person pose estimation with local joint-to-person associations.* In European Conference on Computer Vision, pages 627–642. Springer, 2016. (Cited on page 12.)

[Isola 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou and Alexei A. Efros. *Image-to-Image Translation with Conditional Adversarial Networks.* In CVPR, 2017. (Cited on pages 69 and 75.)

[Jackson 2017] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou and Georgios Tzimiropoulos. *Large pose 3D face reconstruction from a single image via direct volumetric CNN regression.* In 2017 IEEE International Conference on Computer Vision (ICCV), pages 1031–1039. IEEE, 2017. (Cited on page 18.)

[Jaderberg 2015] Max Jaderberg, Karen Simonyan, Andrew Zisserman*et al. Spatial transformer networks.* In Advances in neural information processing systems, pages 2017–2025, 2015. (Cited on pages 4 and 72.)

[Jain 1996] Anil K. Jain, Yu Zhong and Sridhar Lakshmanan. *Object matching using deformable templates.* IEEE Transactions on pattern analysis and machine intelligence, vol. 18, no. 3, pages 267–278, 1996. (Cited on page 16.)

[Joachims 2009] Thorsten Joachims, Thomas Finley and Chun-Nam John Yu. *Cutting-plane training of structural SVMs.* Machine Learning, vol. 77, no. 1, pages 27–59, 2009. (Cited on page 45.)

[Johnson 2010] Sam Johnson and Mark Everingham. *Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation.* In BMVC, volume 2, page 5, 2010. (Cited on page 48.)

[Johnson 2011] Sam Johnson and Mark Everingham. *Learning effective human pose estimation from inaccurate annotation.* In Computer vision and pattern recognition (CVPR), 2011 IEEE conference on, pages 1465–1472. IEEE, 2011. (Cited on page 12.)

[Johnson 2016] Justin Johnson, Alexandre Alahi and Fei-Fei Li. *Perceptual Losses for Real-Time Style Transfer and Super-Resolution.* In ECCV, 2016. (Cited on pages 72 and 75.)

[Jones 1998] Michael J Jones and Tomaso Poggio. *Multidimensional morphable models: A framework for representing and matching object classes.* International Journal of Computer Vision, vol. 29, no. 2, pages 107–131, 1998. (Cited on page 16.)

[Joo 2018] Hanbyul Joo, Tomas Simon and Yaser Sheikh. *Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8320–8329, 2018. (Cited on page 20.)

[Jordan 1994] Michael I. Jordan and Robert A. Jacobs. *Hierarchical Mixtures of Experts and the EM Algorithm.* Neural Computation, vol. 6, no. 2, 1994. (Cited on pages 28 and 29.)

[Jourabloo 2016] Amin Jourabloo and Xiaoming Liu. *Large-pose face alignment via CNN-based dense 3D model fitting.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4188–4196, 2016. (Cited on pages 18 and 26.)

[Ju 1996] Shanon X Ju, Michael J Black and Yaser Yacoob. *Cardboard people: A parameterized model of articulated image motion.* In fg, page 38. IEEE, 1996. (Cited on page 18.)

[Kakadiaris 2000] L Kakadiaris and Dimitris Metaxas. *Model-based estimation of 3D human motion.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pages 1453–1459, 2000. (Cited on page 18.)

[Kanazawa 2018a] Angjoo Kanazawa, Michael J Black, David W Jacobs and Jitendra Malik. *End-to-end recovery of human shape and pose.* In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. (Cited on pages 20 and 69.)

[Kanazawa 2018b] Angjoo Kanazawa, Michael J Black, David W Jacobs and Jitendra Malik. *End-to-end recovery of human shape and pose.* In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. (Cited on page 83.)

[Kanazawa 2018c] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros and Jitendra Malik. *Learning Category-Specific Mesh Reconstruction from Image Collections.* arXiv preprint arXiv:1803.07549, 2018. (Cited on pages 16 and 85.)

[Karras 2017] Tero Karras, Timo Aila, Samuli Laine and Jaakko Lehtinen. *Progressive growing of gans for improved quality, stability, and variation.* arXiv preprint arXiv:1710.10196, 2017. (Cited on pages 69, 77, 81 and 84.)

[Kato 2018] Hiroharu Kato, Yoshitaka Ushiku and Tatsuya Harada. *Neural 3d mesh renderer.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3907–3916, 2018. (Cited on page 85.)

[Kim 2012] Vladimir G. Kim, Wilmot Li, Niloy Mitra, Stephen DiVerdi and Thomas Funkhouser. *Exploring Collections of 3D Models using Fuzzy Correspondences.* Transactions on Graphics (Proc. of SIGGRAPH 2012), vol. 31, no. 4, August 2012. (Cited on page 84.)

[Kim 2013] Jaechul Kim, Ce Liu, Fei Sha and Kristen Grauman. *Deformable spatial pyramid matching for fast dense correspondences.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2307–2314, 2013. (Cited on page 16.)

[Kim 2017a] Hyeongwoo Kim, Michael Zollhöfer, Ayush Tewari, Justus Thies, Christian Richardt and Christian Theobalt. *Inversefacenet: Deep single-shot inverse face rendering from a single image.* arXiv preprint arXiv:1703.10956, 2017. (Cited on page 18.)

[Kim 2017b] Seungryong Kim, Dongbo Min, Bumsub Ham, Sangryul Jeon, Stephen Lin and Kwanghoon Sohn. *Fcss: Fully convolutional self-similarity for dense semantic correspondence.* In Proc. IEEE Conf. Comp. Vision Patt. Recog, volume 1, page 8, 2017. (Cited on page 16.)

[Kinauer 2016] Stefan Kinauer, Maxim Berman and Iasonas Kokkinos. *Monocular Surface Reconstruction Using 3D Deformable Part Models.* In Computer Vision–ECCV 2016 Workshops, pages 296–308. Springer, 2016. (Cited on page 42.)

[King 2015] Davis E King. *Max-margin object detection.* arXiv preprint arXiv:1502.00046, 2015. (Cited on page 34.)

[Kirillov 2017] Alexander Kirillov, Evgeny Levinkov, Bjoern Andres, Bogdan Savchynskyy and Carsten Rother. *Instancecut: from edges to instances with multicut.* In CVPR, volume 3, page 9, 2017. (Cited on page 11.)

[Kocabas 2018] Muhammed Kocabas, Salih Karagoz and Emre Akbas. *Multi-PoseNet: Fast multi-person pose estimation using pose residual network.* In European Conference on Computer Vision, pages 437–453. Springer, 2018. (Cited on page 12.)

[Kokkinos 2007] Iasonas Kokkinos and Alan Yuille. *Unsupervised learning of object deformation models.* In Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, pages 1–8. IEEE, 2007. (Cited on page 16.)

[Komodakis 2007] Nikos Komodakis, Nikos Paragios and Georgios Tziritas. *MRF optimization via dual decomposition: Message-passing revisited.* In Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, pages 1–8. IEEE, 2007. (Cited on page 43.)

[Krizhevsky 2012] Alex Krizhevsky, Ilya Sutskever and Geoffrey E Hinton. *Imagenet classification with deep convolutional neural networks.* In Advances in neural information processing systems, pages 1097–1105, 2012. (Cited on page 8.)

[Kry 2002] Paul G Kry, Doug L James and Dinesh K Pai. *Eigenskin: real time large deformation character skinning in hardware.* In Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation, pages 153–159. ACM, 2002. (Cited on page 19.)

[Lampert 2009] Christoph H Lampert, Matthew B Blaschko and Thomas Hofmann. *Efficient subwindow search: A branch and bound framework for object localization.* IEEE transactions on pattern analysis and machine intelligence, vol. 31, no. 12, page 2129, 2009. (Cited on page 9.)

[Lample 2017] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer and Marc'Aurelio Ranzato. *Fader Networks: Manipulating Images by Sliding Attributes.* In NIPS, 2017. (Cited on page 69.)

[Lassner 2017a] Christoph Lassner, Gerard Pons-Moll and Peter V. Gehler. *A Generative Model of People in Clothing.* In ICCV, 2017. (Cited on pages 69, 72 and 84.)

[Lassner 2017b] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black and Peter V Gehler. *Unite the people: Closing the loop between 3d and 2d human representations.* In IEEE Conf. on Computer

Vision and Pattern Recognition (CVPR), volume 2, page 3, 2017. (Cited on pages 20, 53, 61, 63 and 69.)

[Le 2012] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev and Thomas S Huang. *Interactive facial feature localization*. In European Conference on Computer Vision. Springer, 2012. (Cited on page 31.)

[Learned-Miller 2006] Erik G Learned-Miller. *Data driven image models through continuous joint alignment*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 2, pages 236–250, 2006. (Cited on page 16.)

[LeCun 1998] Yann LeCun, Léon Bottou, Yoshua Bengio and Patrick Haffner. *Gradient-based learning applied to document recognition*. Proceedings of the IEEE, vol. 86, no. 11, pages 2278–2324, 1998. (Cited on page 8.)

[Lee 2015] Chen-Yu Lee, Saining Xie, Patrick W. Gallagher, Zhengyou Zhang and Zhuowen Tu. *Deeply-Supervised Nets*. In AISTATS, 2015. (Cited on pages 44 and 59.)

[Lewis 2000] John P Lewis, Matt Cordner and Nickson Fong. *Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation*. In Proceedings of the 27th annual conference on Computer graphics and interactive techniques, pages 165–172. ACM Press/Addison-Wesley Publishing Co., 2000. (Cited on page 19.)

[Li 2014] Sijin Li and Antoni B Chan. *3d human pose estimation from monocular images with deep convolutional neural network*. In Asian Conference on Computer Vision, pages 332–347. Springer, 2014. (Cited on page 14.)

[Li 2015] Sijin Li, Weichen Zhang and Antoni B Chan. *Maximum-margin structured learning with deep networks for 3d human pose estimation*. In Proceedings of the IEEE International Conference on Computer Vision, pages 2848–2856, 2015. (Cited on page 14.)

[Li 2017] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji and Yichen Wei. *Fully Convolutional Instance-Aware Semantic Segmentation*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2359–2367, 2017. (Cited on page 10.)

[Liang 2015] Xiaodan Liang, Yunchao Wei, Xiaohui Shen, Jianchao Yang, Liang Lin and Shuicheng Yan. *Proposal-free network for instance-level object segmentation*. arXiv preprint arXiv:1509.02636, 2015. (Cited on page 10.)

[Lin 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár and C Lawrence Zitnick. *Microsoft coco: Common objects in context*. In European conference on computer vision, pages 740–755. Springer, 2014. (Cited on pages 10, 14, 31, 53, 55 and 57.)

[Lin 2017] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan and Serge J Belongie. *Feature Pyramid Networks for Object Detection.* In CVPR, volume 1, page 4, 2017. (Cited on pages 9 and 58.)

[Liu 2016a] K.-H. Liu, T.-Y. Chen and C.-S. Chen. *A dataset for view-invariant clothing retrieval and attribute prediction.* In ICMR, 2016. (Cited on page 76.)

[Liu 2016b] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu and Alexander C Berg. *Ssd: Single shot multibox detector.* In European conference on computer vision, pages 21–37. Springer, 2016. (Cited on pages 9 and 77.)

[Liu 2016c] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang and Xiaoou Tang. *Deepfashion: Powering robust clothes recognition and retrieval with rich annotations.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1096–1104, 2016. (Cited on pages 70, 75, 77 and 84.)

[Liu 2017] Shu Liu, Jiaya Jia, Sanja Fidler and Raquel Urtasun. *Sgn: Sequential grouping networks for instance segmentation.* In The IEEE International Conference on Computer Vision (ICCV), 2017. (Cited on page 11.)

[Long 2015] Jonathan Long, Evan Shelhamer and Trevor Darrell. *Fully convolutional networks for semantic segmentation.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3431–3440, 2015. (Cited on pages 10 and 60.)

[Loper 2015] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll and Michael J. Black. *SMPL: A Skinned Multi-Person Linear Model.* ACM Trans. Graphics (Proc. SIGGRAPH Asia), vol. 34, no. 6, pages 248:1–248:16, October 2015. (Cited on pages 18, 19, 20, 52, 55, 61 and 69.)

[Lowe 2004] David G Lowe. *Distinctive image features from scale-invariant keypoints.* International journal of computer vision, vol. 60, no. 2, pages 91–110, 2004. (Cited on page 8.)

[Luvizon 2018] Diogo C Luvizon, David Picard and Hedi Tabia. *2d/3d pose estimation and action recognition using multitask deep learning.* In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), volume 2, 2018. (Cited on page 85.)

[Ma 2017] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars and Luc Van Gool. *Pose Guided Person Image Generation.* In NIPS, 2017. (Cited on pages 69, 72, 77, 79 and 84.)

[Ma 2018] Liqian Ma, Qianru Sun, Stamatious Georgoulis, Luc Van Gool, Bernt Schiele and Mario Fritz. *Disentangled Person Image Generation.* In CVPR, 2018. (Cited on page 77.)

[Mao 2017] Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang and Stephen Paul Smolley. *Least Squares Generative Adversarial Networks*. In ICCV, 2017. (Cited on page 75.)

[Marr 1978] David Marr and Herbert Keith Nishihara. *Representation and recognition of the spatial organization of three-dimensional shapes*. Proc. R. Soc. Lond. B, vol. 200, no. 1140, pages 269–294, 1978. (Cited on pages 18 and 19.)

[Marr 1982] David Marr. Vision: A computational investigation into the human representation and processing of visual information. Henry Holt and Co., Inc., New York, NY, USA, 1982. (Cited on page 9.)

[Martinez 2016] Brais Martinez and Michel F Valstar. *L 2, 1-based regression and prediction accumulation across views for robust facial landmark detection*. Image and Vision Computing, vol. 47, 2016. (Cited on page 34.)

[Martinez 2017] Julieta Martinez, Rayat Hossain, Javier Romero and James J Little. *A simple yet effective baseline for 3d human pose estimation*. In International Conference on Computer Vision, volume 1, page 5, 2017. (Cited on page 13.)

[Martins 2011] André FT Martins, Noah A Smith, Pedro MQ Aguiar and Mário AT Figueiredo. *Dual decomposition with many overlapping components*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 238–249. Association for Computational Linguistics, 2011. (Cited on page 42.)

[Mathias 2014] Markus Mathias, Rodrigo Benenson, Marco Pedersoli and Luc Van Gool. *Face detection without bells and whistles*. In European Conference on Computer Vision. Springer, 2014. (Cited on pages 30 and 34.)

[Matthews 2004] Iain Matthews and Simon Baker. *Active appearance models revisited*. International journal of computer vision, vol. 60, no. 2, pages 135–164, 2004. (Cited on page 16.)

[Mehta 2017] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu and Christian Theobalt. *Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision*. In 3D Vision (3DV), 2017 Fifth International Conference on, 2017. (Cited on page 13.)

[Mikolajczyk 2005] Krystian Mikolajczyk and Cordelia Schmid. *A performance evaluation of local descriptors*. IEEE transactions on pattern analysis and machine intelligence, vol. 27, no. 10, pages 1615–1630, 2005. (Cited on page 8.)

[Mitra 2013] Niloy Mitra, Michael Wand, Hao Richard Zhang, Daniel Cohen-Or, Vladimir Kim and Qi-Xing Huang. *Structure-aware shape processing.* In SIGGRAPH Asia 2013 Courses, page 1. ACM, 2013. (Cited on page 84.)

[Mo 2018] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas and Hao Su. *PartNet: A Large-scale Benchmark for Fine-grained and Hierarchical Part-level 3D Object Understanding.* arXiv preprint arXiv:1812.02713, 2018. (Cited on page 85.)

[MoCap 2003] CMU MoCap. *mocap. cs. cmu. edu*, 2003. (Cited on page 61.)

[Mori 2002] Greg Mori and Jitendra Malik. *Estimating human body configurations using shape context matching.* In European conference on computer vision, pages 666–680. Springer, 2002. (Cited on page 13.)

[Neverova 2017] Natalia Neverova, Christian Wolf, Florian Nebout and Graham Taylor. *Hand pose estimation through weakly-supervised learning of a rich intermediate representation.* Computer Vision and Image Understanding, 2017. (Cited on page 61.)

[Newell 2016] Alejandro Newell, Kaiyu Yang and Jia Deng. *Stacked hourglass networks for human pose estimation.* In European Conference on Computer Vision, pages 483–499. Springer, 2016. (Cited on pages 12, 27, 30, 40 and 59.)

[Newell 2017] Alejandro Newell, Zhiao Huang and Jia Deng. *Associative embedding: End-to-end learning for joint detection and grouping.* In Advances in Neural Information Processing Systems, pages 2277–2287, 2017. (Cited on pages 11 and 12.)

[Omran 2018] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler and Bernt Schiele. *Neural body fitting: Unifying deep learning and model based human pose and shape estimation.* In 2018 International Conference on 3D Vision (3DV), pages 484–494. IEEE, 2018. (Cited on page 20.)

[O'rourke 1980] Joseph O'rourke and Norman I Badler. *Model-based image analysis of human motion using constraint propagation.* IEEE Transactions on Pattern Analysis and Machine Intelligence, no. 6, pages 522–536, 1980. (Cited on page 13.)

[Papandreou 2008] George Papandreou and Petros Maragos. *Adaptive and constrained algorithms for inverse compositional active appearance model fitting.* In Computer Vision and Pattern Recognition, 2008. IEEE Conference on. IEEE, 2008. (Cited on page 34.)

[Papandreou 2015] George Papandreou, Iasonas Kokkinos and Pierre-André Savalle. *Modeling local and global deformations in Deep Learning: Epitomic convolution, Multiple Instance Learning, and sliding window detection.*

In IEEE Conference on Computer Vision and Pattern Recognition, 2015, Boston, MA, USA, June 7-12, 2015, 2015. (Cited on page 3.)

[Papandreou 2017] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler and Kevin Murphy. *Towards accurate multi-person pose estimation in the wild.* In CVPR, volume 3, page 6, 2017. (Cited on pages 12 and 41.)

[Papandreou 2018] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson and Kevin Murphy. *PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model.* arXiv preprint arXiv:1803.08225, 2018. (Cited on pages 11 and 12.)

[Parameswaran 2004] Vasu Parameswaran and Rama Chellappa. *View independent human body pose estimation from a single perspective image.* In Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, volume 2, pages II–II. IEEE, 2004. (Cited on page 13.)

[Patel 2009] Ankur Patel and William AP Smith. *3d morphable face models revisited.* In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 1327–1334. IEEE, 2009. (Cited on page 17.)

[Pavlakos 2017] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis and Kostas Daniilidis. *Coarse-to-fine volumetric prediction for single-image 3D human pose.* In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, pages 1263–1272. IEEE, 2017. (Cited on pages 13, 14, 39, 46 and 48.)

[Pavlakos 2018] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou and Kostas Daniilidis. *Learning to Estimate 3D Human Pose and Shape from a Single Color Image.* arXiv preprint arXiv:1805.04092, 2018. (Cited on page 20.)

[Paysan 2009] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani and Thomas Vetter. *A 3D face model for pose and illumination invariant face recognition.* In AVSS. IEEE, 2009. (Cited on pages 17, 26 and 30.)

[Pepik 2015] Bojan Pepik, Michael Stark, Peter V. Gehler and Bernt Schiele. *Multi-View and 3D Deformable Part Models.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 11, pages 2232–2245, 2015. (Cited on page 45.)

[Pinheiro 2015] Pedro O Pinheiro, Ronan Collobert and Piotr Dollár. *Learning to segment object candidates.* In Advances in Neural Information Processing Systems, pages 1990–1998, 2015. (Cited on page 10.)

[Pinheiro 2016] Pedro O Pinheiro, Tsung-Yi Lin, Ronan Collobert and Piotr Dollár. *Learning to refine object segments.* In European Conference on Computer Vision, pages 75–91. Springer, 2016. (Cited on page 10.)

[Pishchulin 2011] Leonid Pishchulin, Arjun Jain, Christian Wojek, Mykhaylo Andriluka, Thorsten Thormählen and Bernt Schiele. *Learning people detection models from few training samples.* In CVPR, 2011. (Cited on page 61.)

[Pishchulin 2012] Leonid Pishchulin, Arjun Jain, Mykhaylo Andriluka, Thorsten Thormählen and Bernt Schiele. *Articulated people detection and pose estimation: Reshaping the future.* In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 3178–3185. IEEE, 2012. (Cited on pages 12 and 61.)

[Pishchulin 2013] Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler and Bernt Schiele. *Poselet conditioned pictorial structures.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 588–595, 2013. (Cited on page 12.)

[Pishchulin 2016] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler and Bernt Schiele. *Deepcut: Joint subset partition and labeling for multi person pose estimation.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4929–4937, 2016. (Cited on page 12.)

[Pons-Moll 2014] Gerard Pons-Moll, David J Fleet and Bodo Rosenhahn. *Posebits for monocular human pose estimation.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2337–2344, 2014. (Cited on page 14.)

[Pons-Moll 2015a] Gerard Pons-Moll, Javier Romero, Naureen Mahmood and Michael J. Black. *Dyna: A Model of Dynamic Human Shape in Motion.* ACM Transactions on Graphics, (Proc. SIGGRAPH), vol. 34, no. 4, August 2015. (Cited on page 20.)

[Pons-Moll 2015b] Gerard Pons-Moll, Jonathan Taylor, Jamie Shotton, Aaron Hertzmann and Andrew Fitzgibbon. *Metric regression forests for correspondence estimation.* International Journal of Computer Vision, vol. 113, no. 3, pages 163–175, 2015. (Cited on page 52.)

[Rajamanoharan 2015] Georgia Rajamanoharan and Timothy F Cootes. *Multi-View Constrained Local Models for Large Head Angle Facial Tracking.* In Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015. (Cited on page 35.)

[Ramakrishna 2012] Varun Ramakrishna, Takeo Kanade and Yaser Sheikh. *Reconstructing 3d human pose from 2d image landmarks.* In European conference on computer vision, pages 573–586. Springer, 2012. (Cited on page 13.)

[Ramakrishna 2014] Varun Ramakrishna, Daniel Munoz, Martial Hebert, James Andrew Bagnell and Yaser Sheikh. *Pose machines: Articulated pose estimation via inference machines.* In European Conference on Computer Vision, pages 33–47. Springer, 2014. (Cited on page 12.)

[Redmon 2016] Joseph Redmon, Santosh Divvala, Ross Girshick and Ali Farhadi. *You only look once: Unified, real-time object detection.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016. (Cited on page 9.)

[Ren 2015] Shaoqing Ren, Kaiming He, Ross Girshick and Jian Sun. *Faster r-cnn: Towards real-time object detection with region proposal networks.* In Advances in neural information processing systems, pages 91–99, 2015. (Cited on pages 6, 9, 10, 41 and 58.)

[Richardson 2016] Elad Richardson, Matan Sela and Ron Kimmel. *3D face reconstruction by learning from synthetic data.* In 3D Vision (3DV), 2016 Fourth International Conference on, pages 460–469. IEEE, 2016. (Cited on page 18.)

[Robinette 1999] Kathleen M Robinette, Hans Daanen and Eric Paquet. *The CAESAR project: a 3-D surface anthropometry survey.* In 3-D Digital Imaging and Modeling, 1999. Proceedings. Second International Conference on, pages 380–386. IEEE, 1999. (Cited on page 19.)

[Rocco 2017] Ignacio Rocco, Relja Arandjelovic and Josef Sivic. *Convolutional neural network architecture for geometric matching.* In Proc. CVPR, volume 2, 2017. (Cited on page 16.)

[Rocco 2018] Ignacio Rocco, Relja Arandjelovic and Josef Sivic. *End-to-end weakly-supervised semantic alignment.* In Proc. CVPR, 2018. (Cited on page 16.)

[Rogez 2016] Grégory Rogez and Cordelia Schmid. *Mocap-guided data augmentation for 3d pose estimation in the wild.* In Advances in Neural Information Processing Systems, pages 3108–3116, 2016. (Cited on pages 14, 48 and 61.)

[Rogez 2017] Gregory Rogez, Philippe Weinzaepfel and Cordelia Schmid. *Lcr-net: Localization-classification-regression for human pose.* In CVPR 2017-IEEE Conference on Computer Vision & Pattern Recognition, 2017. (Cited on page 14.)

[Rohr 1994] Karl Rohr. *Towards model-based recognition of human movements in image sequences.* CVGIP-Image Understanding, vol. 59, no. 1, pages 94–115, 1994. (Cited on page 18.)

[Romdhani 2005] Sami Romdhani and Thomas Vetter. *Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior.* In Computer Vision and Pattern Recognition, 2005. CVPR

2005. IEEE Computer Society Conference on, volume 2, pages 986–993. IEEE, 2005. (Cited on pages 18 and 26.)

[Ronchi 2017] Matteo Ruggero Ronchi and Pietro Perona. *Benchmarking and Error Diagnosis in Multi-Instance Pose Estimation.* In The IEEE International Conference on Computer Vision (ICCV), Oct 2017. (Cited on pages 55 and 57.)

[Rowley 1998] Henry A Rowley, Shumeet Baluja and Takeo Kanade. *Neural network-based face detection.* IEEE Transactions on pattern analysis and machine intelligence, vol. 20, no. 1, pages 23–38, 1998. (Cited on page 9.)

[Sagonas 2013] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou and Maja Pantic. *300 faces in-the-wild challenge: The first facial landmark localization challenge.* In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 397–403, 2013. (Cited on page 34.)

[Sagonas 2016] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou and Maja Pantic. *300 faces in-the-wild challenge: Database and results.* Image and Vision Computing, vol. 47, 2016. (Cited on page 34.)

[Salimans 2016] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford and X. Chen. *Improved techniques for training gans.* In NIPS, 2016. (Cited on page 77.)

[Sapp 2010a] Benjamin Sapp, Chris Jordan and Ben Taskar. *Adaptive pose priors for pictorial structures.* In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pages 422–429. IEEE, 2010. (Cited on page 12.)

[Sapp 2010b] Benjamin Sapp, Alexander Toshev and Ben Taskar. *Cascaded models for articulated pose estimation.* Computer Vision–ECCV 2010, pages 406–420, 2010. (Cited on page 42.)

[Sapp 2013] Ben Sapp and Ben Taskar. *Modec: Multimodal decomposable models for human pose estimation.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3674–3681, 2013. (Not cited.)

[Schönborn 2017] Sandro Schönborn, Bernhard Egger, Andreas Morel-Forster and Thomas Vetter. *Markov chain monte carlo for automated face image analysis.* International Journal of Computer Vision, vol. 123, no. 2, pages 160–183, 2017. (Cited on page 18.)

[Sela 2017] Matan Sela, Elad Richardson and Ron Kimmel. *Unrestricted facial geometry reconstruction using image-to-image translation.* In Computer Vision

(ICCV), 2017 IEEE International Conference on, pages 1585–1594. IEEE, 2017. (Cited on page 18.)

[Seo 2003] Hyewon Seo, Frederic Cordier and Nadia Magnenat-Thalmann. *Synthesizing animatable body models with parameterized shape modifications.* In Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation, pages 120–125. Eurographics Association, 2003. (Cited on page 19.)

[Sermanet 2013a] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus and Yann LeCun. *Overfeat: Integrated recognition, localization and detection using convolutional networks.* arXiv preprint arXiv:1312.6229, 2013. (Cited on page 9.)

[Sermanet 2013b] Pierre Sermanet, Koray Kavukcuoglu, Soumith Chintala and Yann LeCun. *Pedestrian detection with unsupervised multi-stage feature learning.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3626–3633, 2013. (Cited on page 9.)

[Shakhnarovich 2003] Gregory Shakhnarovich, Paul Viola and Trevor Darrell. *Fast pose estimation with parameter-sensitive hashing.* In null, page 750. IEEE, 2003. (Cited on page 13.)

[Shen 2015] J. Shen, S. Zafeiriou, G. Chrysos, J. Kossaifi, G. Tzimiropoulos and M. Pantic. *The First Facial Landmark Tracking in-the-Wild Challenge: Benchmark and Results.* In Proceedings of IEEE International Conference on Computer Vision, 300 Videos in the Wild (300-VW): Facial Landmark Tracking in-the-Wild Challenge & Workshop (ICCVW'15), December 2015. (Cited on page 35.)

[Shotton 2008] Jamie Shotton, Matthew Johnson and Roberto Cipolla. *Semantic texton forests for image categorization and segmentation.* In Computer vision and pattern recognition, 2008. CVPR 2008. IEEE Conference on, pages 1–8. IEEE, 2008. (Cited on pages 8 and 10.)

[Shu 2017] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman and Dimitris Samaras. *Neural Face Editing with Intrinsic Image Disentangling.* In CVPR, 2017. (Cited on page 69.)

[Shu 2018] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios and Iasonas Kokkinos. *Deforming autoencoders: Unsupervised disentangling of shape and appearance.* In Proceedings of the European Conference on Computer Vision (ECCV), pages 650–665, 2018. (Cited on page 85.)

[Siarohin 2018] Aliaksandr Siarohin, Enver Sangineto, Stephane Lathuiliere and Nicu Sebe. *Deformable GANs for Pose-based Human Image Generation.* In CVPR, 2018. (Cited on pages 69, 76, 77, 78, 79 and 84.)

[Sidenbladh 2000] Hedvig Sidenbladh, Michael J Black and David J Fleet. *Stochastic tracking of 3D human figures using 2D image motion.* In European conference on computer vision, pages 702–718. Springer, 2000. (Cited on page 18.)

[Sidi 2011] Oana Sidi, Oliver van Kaick, Yanir Kleiman, Hao Zhang and Daniel Cohen-Or. Unsupervised co-segmentation of a set of shapes via descriptor-space spectral clustering, volume 30. ACM, 2011. (Cited on page 84.)

[Sigal 2004] Leonid Sigal, Michael Isard, Benjamin H Sigelman and Michael J Black. *Attractive people: Assembling loose-limbed models using non-parametric belief propagation.* In Advances in neural information processing systems, pages 1539–1546, 2004. (Cited on page 18.)

[Sigal 2010] Leonid Sigal, Alexandru O Balan and Michael J Black. *Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion.* International journal of computer vision, vol. 87, no. 1-2, page 4, 2010. (Cited on page 13.)

[Sigal 2012] Leonid Sigal, Michael Isard, Horst Haussecker and Michael J Black. *Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation.* International journal of computer vision, vol. 98, no. 1, pages 15–48, 2012. (Cited on page 14.)

[Simo-Serra 2012] Edgar Simo-Serra, Arnau Ramisa, Guillem Alenyà, Carme Torras and Francesc Moreno-Noguer. *Single image 3D human pose estimation from noisy observations.* In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 2673–2680. IEEE, 2012. (Cited on page 13.)

[Simo-Serra 2013] Edgar Simo-Serra, Ariadna Quattoni, Carme Torras and Francesc Moreno-Noguer. *A joint model for 2d and 3d pose estimation from a single image.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3634–3641, 2013. (Cited on page 13.)

[Simonyan 2014a] Karen Simonyan and Andrew Zisserman. *Two-stream convolutional networks for action recognition in videos.* In Advances in neural information processing systems, pages 568–576, 2014. (Cited on page 85.)

[Simonyan 2014b] Karen Simonyan and Andrew Zisserman. *Very deep convolutional networks for large-scale image recognition.* arXiv preprint arXiv:1409.1556, 2014. (Cited on pages 8 and 74.)

[Sirovich 1987] Lawrence Sirovich and Michael Kirby. *Low-dimensional procedure for the characterization of human faces.* Josa a, vol. 4, no. 3, pages 519–524, 1987. (Cited on page 16.)

[Sminchisescu 2003] Cristian Sminchisescu and Bill Triggs. *Estimating articulated human motion with covariance scaled sampling.* The International Journal of Robotics Research, vol. 22, no. 6, pages 371–391, 2003. (Cited on page 18.)

[Sminchisescu 2005] Cristian Sminchisescu, Atul Kanaujia, Zhiguo Li and Dimitris Metaxas. *Discriminative density propagation for 3d human motion estimation.* In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 1, pages 390–397. IEEE, 2005. (Cited on page 13.)

[Staib 1992] Lawrence H Staib and James S Duncan. *+.* IEEE Transactions on Pattern Analysis & Machine Intelligence, no. 11, pages 1061–1075, 1992. (Cited on page 15.)

[Sun 2017] Xiao Sun, Bin Xiao, Shuang Liang and Yichen Wei. *Integral human pose regression.* arXiv preprint arXiv:1711.08229, 2017. (Cited on page 14.)

[Sun 2018] Xiao Sun, Jiaxiang Shang, Shuang Liang and Yichen Wei. *Compositional human pose regression.* In ECCV, 2018. (Cited on pages 14, 46 and 48.)

[Taylor 2012] Jonathan Taylor, Jamie Shotton, Toby Sharp and Andrew W. Fitzgibbon. *The Vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation.* In CVPR, 2012. (Cited on page 52.)

[Tekin 2016] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit and Pascal Fua. *Structured prediction of 3d human pose with deep neural networks.* arXiv preprint arXiv:1605.05180, 2016. (Cited on pages 14 and 40.)

[Tekin 2017] Bugra Tekin, Pablo Marquez Neila, Mathieu Salzmann and Pascal Fua. *Learning to fuse 2d and 3d image cues for monocular body pose estimation.* In International Conference on Computer Vision (ICCV), 2017. (Cited on page 14.)

[Thewlis 2017] James Thewlis, Hakan Bilen and Andrea Vedaldi. *Unsupervised learning of object frames by dense equivariant image labelling.* In Advances in Neural Information Processing Systems, pages 844–855, 2017. (Cited on pages 16 and 85.)

[Thompson 1942] Darcy Wentworth Thompson*et al. On growth and form.* On growth and form., 1942. (Cited on page 15.)

[Tome 2017] Denis Tome, Christopher Russell and Lourdes Agapito. *Lifting from the deep: Convolutional 3d pose estimation from a single image.* CVPR 2017 Proceedings, pages 2500–2509, 2017. (Cited on pages 14, 39 and 48.)

[Tompson 2014] Jonathan J Tompson, Arjun Jain, Yann LeCun and Christoph Bregler. *Joint training of a convolutional network and a graphical model for human pose estimation.* In Advances in neural information processing systems, pages 1799–1807, 2014. (Cited on pages 12 and 40.)

[Toshev 2014] Alexander Toshev and Christian Szegedy. *Deeppose: Human pose estimation via deep neural networks.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1653–1660, 2014. (Cited on pages 12 and 14.)

[Tran 2017] Anh Tuan Tran, Tal Hassner, Iacopo Masi and Gérard Medioni. *Regressing robust and discriminative 3D morphable models with a very deep neural network.* In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, pages 1493–1502. IEEE, 2017. (Cited on page 18.)

[Trigeorgis 2016] George Trigeorgis, Patrick Snape, Mihalis A Nicolaou, Epameinondas Antonakos and Stefanos Zafeiriou. *Mnemonic Descent Method: A recurrent process applied for end-to-end face alignment.* In Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition, 2016. (Cited on pages 17 and 34.)

[Tsogkas 2015] Stavros Tsogkas, Iasonas Kokkinos, George Papandreou and Andrea Vedaldi. *Deep Learning for Semantic Part Segmentation with High-Level Guidance.* arXiv preprint arXiv:1505.02438, 2015. (Cited on page 27.)

[Turk 1991] Matthew Turk and Alex Pentland. *Eigenfaces for recognition.* Journal of cognitive neuroscience, vol. 3, no. 1, pages 71–86, 1991. (Cited on page 16.)

[Tzimiropoulos 2014] Georgios Tzimiropoulos and Maja Pantic. *Gauss-newton deformable part models for face alignment in-the-wild.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014. (Cited on page 34.)

[Uhrig 2016] Jonas Uhrig, Marius Cordts, Uwe Franke and Thomas Brox. *Pixel-level encoding and depth layering for instance-level semantic labeling.* In German Conference on Pattern Recognition, pages 14–25. Springer, 2016. (Cited on page 10.)

[Uijlings 2013] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers and Arnold WM Smeulders. *Selective search for object recognition.* International journal of computer vision, vol. 104, no. 2, pages 154–171, 2013. (Cited on page 9.)

[Ulyanov 2017] Dmitry Ulyanov, Andrea Vedaldi and Victor Lempitsky. *Improved Texture Networks: Maximizing Quality and Diversity in Feed-forward Stylization and Texture Synthesis.* In CVPR, 2017. (Cited on page 72.)

[Uricár 2015] Michal Uricár, Vojtech Franc and Václav Hlavác. *Facial Landmark Tracking by Tree-based Deformable Part Model Based Detector.* In Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015. (Cited on page 35.)

[Uřičář 2016] Michal Uřičář, Vojtěch Franc, Diego Thomas, Akihiro Sugimoto and Václav Hlaváč. *Multi-view facial landmark detector learned by the structured output SVM.* Image and Vision Computing, vol. 47, 2016. (Cited on page 34.)

[Vaillant 1994] Régis Vaillant, Christophe Monrocq and Yann Le Cun. *Original approach for the localisation of objects in images.* IEE Proceedings-Vision, Image and Signal Processing, vol. 141, no. 4, pages 245–250, 1994. (Cited on page 9.)

[Varol 2017] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev and Cordelia Schmid. *Learning from Synthetic Humans.* In CVPR, 2017. (Cited on pages 14, 53, 55, 61, 67 and 69.)

[Varol 2018] Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev and Cordelia Schmid. *BodyNet: Volumetric Inference of 3D Human Body Shapes.* In ECCV, 2018. (Cited on page 20.)

[Vetter 1997a] Thomas Vetter, Michael J Jones and Tomaso Poggio. *A bootstrapping algorithm for learning linear models of object classes.* In Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on, pages 40–46. IEEE, 1997. (Cited on page 16.)

[Vetter 1997b] Thomas Vetter, Michael J. Jones and Tomaso A. Poggio. *A bootstrapping algorithm for learning linear models of object classes.* In 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97), June 17-19, 1997, San Juan, Puerto Rico, pages 40–46, 1997. (Cited on page 83.)

[Vetter 1997c] Thomas Vetter and Tomaso Poggio. *Linear object classes and image synthesis from a single example image.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pages 733–742, 1997. (Cited on page 16.)

[Viola 2001] Paul Viola and Michael Jones. *Rapid object detection using a boosted cascade of simple features.* In Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, volume 1, pages I–I. IEEE, 2001. (Cited on page 9.)

[von Marcard 2018] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn and Gerard Pons-Moll. *Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera.* In European Conference on Computer Vision (ECCV), sep 2018. (Cited on page 14.)

[Wang 2003] Z. Wang, Eero P. Simoncelli and Alan C. Bovik. *Multi-scale structural similarity for image quality assessment.* In ACSSC, 2003. (Cited on page 76.)

[Wang 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh and Eero P Simoncelli. *Image quality assessment: from error visibility to structural similarity.* IEEE transactions on image processing, vol. 13, no. 4, pages 600–612, 2004. (Cited on page 76.)

[Wang 2014] Chunyu Wang, Yizhou Wang, Zhouchen Lin, Alan L Yuille and Wen Gao. *Robust estimation of 3d human poses from a single image.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2361–2368, 2014. (Cited on page 13.)

[Wang 2016] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang and Luc Van Gool. *Temporal segment networks: Towards good practices for deep action recognition.* In European Conference on Computer Vision, pages 20–36. Springer, 2016. (Cited on page 85.)

[Wang 2018a] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz and Bryan Catanzaro. *Video-to-video synthesis.* arXiv preprint arXiv:1808.06601, 2018. (Cited on page 84.)

[Wang 2018b] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Kautz Jan and Catanzaro Bryan. *High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs.* In CVPR, 2018. (Cited on pages 69 and 75.)

[Wei 2016a] Lingyu Wei, Qixing Huang, Duygu Ceylan, Etienne Vouga and Hao Li. *Dense Human Body Correspondences Using Convolutional Networks.* In CVPR, 2016. (Cited on page 52.)

[Wei 2016b] Shih-En Wei, Varun Ramakrishna, Takeo Kanade and Yaser Sheikh. *Convolutional pose machines.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4724–4732, 2016. (Cited on pages 12, 40 and 59.)

[Wu 2015] Yue Wu and Qiang Ji. *Shape augmented regression method for face alignment.* In Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015. (Cited on page 35.)

[Xiao 2015] Shengtao Xiao, Shuicheng Yan and Ashraf A Kassim. *Facial landmark detection via progressive initialization.* In Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015. (Cited on page 35.)

[Xiao 2018] Bin Xiao, Haiping Wu and Yichen Wei. *Simple Baselines for Human Pose Estimation and Tracking.* arXiv preprint arXiv:1804.06208, 2018. (Cited on page 12.)

[Xiong 2013] Xuehan Xiong and Fernando De la Torre. *Supervised descent method and its applications to face alignment.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 532–539, 2013. (Cited on page 16.)

[Xu 2017] Kai Xu, Vladimir G Kim, Qixing Huang and Evangelos Kalogerakis. *Data-driven shape analysis and processing.* In Computer Graphics Forum, volume 36, pages 101–132. Wiley Online Library, 2017. (Cited on page 84.)

[Yang 2013] Yi Yang and Deva Ramanan. *Articulated human detection with flexible mixtures of parts.* IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 12, pages 2878–2890, 2013. (Cited on page 12.)

[Yang 2015] Jing Yang, Jiankang Deng, Kaihua Zhang and Qingshan Liu. *Facial shape tracking via spatio-temporal cascade shape regression.* In Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015. (Cited on page 35.)

[Yang 2016] Wei Yang, Wanli Ouyang, Hongsheng Li and Xiaogang Wang. *End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3073–3082, 2016. (Cited on page 40.)

[Yang 2017] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li and Xiaogang Wang. *Learning feature pyramids for human pose estimation.* In The IEEE International Conference on Computer Vision (ICCV), volume 2, 2017. (Cited on page 12.)

[Yasin 2016] Hashim Yasin, Umar Iqbal, Bjorn Kruger, Andreas Weber and Juergen Gall. *A dual-source approach for 3D pose estimation from a single image.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4948–4956, 2016. (Cited on page 48.)

[Yeh 2017] Raymond A. Yeh, Chen Chen, Teck-Yian Lim, Mark Hasegawa-Johnson and Minh N. Do. *Semantic Image Inpainting with Perceptual and Contextual Losses.* In CVPR, 2017. (Cited on page 73.)

[Yi 2016] Li Yi, Vladimir G Kim, Duygu Ceylan, I Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, Leonidas Guibas*et al. A scalable active framework for region annotation in 3d shape collections.* ACM Transactions on Graphics (TOG), vol. 35, page 210, 2016. (Cited on page 84.)

[Yi 2017] Li Yi, Leonidas Guibas, Aaron Hertzmann, Vladimir G Kim, Hao Su and Ersin Yumer. *Learning hierarchical shape segmentation and labeling from online repositories.* arXiv preprint arXiv:1705.01661, 2017. (Cited on page 85.)

[Yuille 1991] Alan L Yuille. *Deformable templates for face recognition.* Journal of Cognitive Neuroscience, vol. 3, no. 1, pages 59–70, 1991. (Cited on page 25.)

[Yuille 1992] Alan L Yuille, Peter W Hallinan and David S Cohen. *Feature extraction from faces using deformable templates.* International journal of computer vision, vol. 8, no. 2, pages 99–111, 1992. (Cited on page 15.)

[Zafeiriou 2017] Stefanos Zafeiriou, George Trigeorgis, Grigorios Chrysos, Jiankang Deng and Jie Shen. *The menpo facial landmark localisation challenge: A step towards the solution.* In CVPR Workshops, 2017. (Cited on page 76.)

[Zanfir 2018] Andrei Zanfir, Elisabeta Marinoiu and Cristian Sminchisescu. *Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes–The Importance of Multiple Scene Constraints.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2148–2157, 2018. (Cited on page 20.)

[Zhang 2015a] Yuting Zhang, Kihyuk Sohn, Ruben Villegas, Gang Pan and Honglak Lee. *Improving object detection with deep convolutional networks via bayesian optimization and structured prediction.* In CVPR, pages 249–258, 2015. (Cited on page 45.)

[Zhang 2015b] Ziyu Zhang, Alexander G Schwing, Sanja Fidler and Raquel Urtasun. *Monocular object instance segmentation and depth ordering with cnns.* In Proceedings of the IEEE International Conference on Computer Vision, pages 2614–2622, 2015. (Cited on page 10.)

[Zhang 2017] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang and Stan Z Li. *S3FD: Single Shot Scale-Invariant Face Detector.* In CVPR, 2017. (Cited on page 76.)

[Zhao 2016] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang and Jiaya Jia. *Pyramid scene parsing network.* arXiv preprint arXiv:1612.01105, 2016. (Cited on page 64.)

[Zhao 2018a] Bo Zhao, Xiao Wu, Zhi-Qi Cheng, Hao Liu and Jiashi Feng. *Multi-View Image Generation from a Single-View.* In ACM on Multimedia Conference, 2018. (Cited on page 77.)

[Zhao 2018b] Ruiqi Zhao, Yan Wang and Aleix M Martinez. *A simple, fast and highly-accurate algorithm to recover 3d shape from 2d landmarks on a single image.* IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 12, pages 3059–3066, 2018. (Cited on page 13.)

[Zhou 2015] Tinghui Zhou, Yong Jae Lee, Stella X Yu and Alyosha A Efros. *Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1191–1200, 2015. (Cited on page 16.)

[Zhou 2016a] Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang and Alexei A Efros. *Learning dense correspondence via 3d-guided cycle consistency.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 117–126, 2016. (Cited on page 16.)

[Zhou 2016b] Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang and Yichen Wei. *Deep kinematic pose regression.* In European Conference on Computer Vision, pages 186–201. Springer, 2016. (Cited on page 14.)

[Zhou 2016c] Yuxiang Zhou, Epameinondas Antonakos, Joan Alabort-i Medina, Anastasios Roussos and Stefanos Zafeiriou. *Estimating Correspondences of Deformable Objects "In-The-Wild".* In The IEEE Conference on Computer Vision and Pattern Recognition, June 2016. (Cited on page 36.)

[Zhou 2017] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos and Kostas Daniilidis. *Sparse representation for 3D shape estimation: A convex relaxation approach.* IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 8, pages 1648–1661, 2017. (Cited on page 13.)

[Zhu 2015] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi and Stan Z Li. *High-fidelity pose and expression normalization for face recognition in the wild.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. (Cited on page 33.)

[Zhu 2016] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi and Stan Z Li. *Face alignment across large poses: A 3d solution.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 146–155, 2016. (Cited on pages 18, 26 and 30.)

[Zolfaghari 2017] Mohammadreza Zolfaghari, Gabriel L Oliveira, Nima Sedaghat and Thomas Brox. *Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection.* In Computer Vision (ICCV), 2017 IEEE International Conference on, pages 2923–2932. IEEE, 2017. (Cited on page 85.)

[Zuffi 2015] Silvia Zuffi and Michael J Black. *The stitched puppet: A graphical model of 3D human shape and pose.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3537–3546, 2015. (Cited on page 20.)

[Zuffi 2017] Silvia Zuffi, Angjoo Kanazawa, David W Jacobs and Michael J Black. *3D Menagerie: Modeling the 3D shape and pose of animals.* In CVPR, pages 5524–5532, 2017. (Cited on page 84.)

[Zuffi 2018] Silvia Zuffi, Angjoo Kanazawa and Michael J Black. *Lions and Tigers and Bears: Capturing Non-Rigid, 3D, Articulated Shape From Images.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3955–3963, 2018. (Cited on page 84.)

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Apprentissage de Correspondances Image-Surface

Mots clés : Vision par ordinateur, Apprentissage automatique, Correspondances Dense, Correspondances Image-Surface

Résumé :

Cette thèse se concentre sur le développement de modèles de représentation dense d'objets 3-D á partir d'images. L'objectif de ce travail est d'améliorer les modèles surfaciques 3-D fournis par les systèmes de vision par ordinateur, en utilisant de nouveaux éléments tirés des images, plutôt que les annotations habituellement utilisées, ou que les modèles basés sur une division de l'objet en différents parties. Des réseaux neuronaux convolutifs (CNNs) sont utilisés pour associer de manière dense les pixels d'une image avec les coordonnées 3-D d'un modèle de l'objet considéré. Cette méthode permet de résoudre très simplement une multitude de tâches de vision par ordinateur, telles que le transfert d'apparence, la localisation de repères ou la segmentation sémantique, en utilisant la correspondance entre une solution sur le modèle surfacique 3-D et l'image 2-D considérée. On démontre qu'une correspondance géométrique entre un modèle 3-D et une image peut être établie pour le visage et le corps humains.

Title : Learning Image-to-Surface Correspondence

Keywords : Computer Vision, Machine Learning, Dense Correspondence, Image-to-Surface Correspondence

Abstract :

This thesis addresses the task of establishing a dense correspondence between an image and a 3D object template. We aim to bring vision systems closer to a surface-based 3D understanding of objects by extracting information that is complementary to existing landmark- or part-based representations. We use convolutional neural networks (CNNs) to densely associate pixels with intrinsic coordinates of 3D object templates. Through the established correspondences we effortlessly solve a multitude of visual tasks, such as appearance transfer, landmark localization and semantic segmentation by transferring solutions from the template to an image. We show that geometric correspondence between an image and a 3D model can be effectively inferred for both the human face and the human body.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France