Università
degli Studi
di Padova

Dipartimento
di Ingegneria
dell'Informazione

Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Corso di Laurea Magistrale in
Computer Engineering

Multi-camera Multiple Human Parsing: Instance-level Human Body Parts Segmentation from Multi-view Images

Supervisor:
Prof. Stefano Ghidoni

Co-supervisors:
Prof. Matteo Terreran
Dr. Leonardo Barcellona

Student:
Laura Bragagnolo

ID number:
2026969

Academic year 2022/2023

Abstract

Multi-human parsing is an important and challenging task in vision-based human understanding, grouping together human body parts segmentation and human instance segmentation.

Although recent deep-learning-based techniques achieve notable results on multi-human parsing datasets, many challenges still remain unresolved. One of them consists in accurately segmenting human bodies in images in which people are very close to each other or overlap. In such cases, multi-human parsing techniques struggle to properly segment human instances and to associate detected body parts to the correct person. This is confirmed by an in-depth analysis provided in this thesis on current state-of-the-art networks for multi-human parsing, which highlights significant issues in presence of severe occlusions between people in the image. To solve this problem, this thesis proposes to exploit multi-view information, based on the intuition that people occluded in an image taken from a particular point of view could be easily separated if framed from a different angle.

Motivated by the absence of a suitable multi-view dataset in the literature, this work proposes to exploit the human instance segmentation task to improve multi-human parsing on strong occlusions. A novel learning framework is introduced to take advantage of human instance segmentation as auxiliary information to guide the multi-human parsing task. Network learning is driven by human segmentation loss functions evaluated on single-views, aiming at improving foreground human instance discrimination, and multi-view instance and body parts prediction consistency, to impose coherent instance and semantic predictions across multiple views of the same scene. The multi-view loss term exploits 3D knowledge to separate overlapping bodies and to provide sparse supervision to human parsing.

To validate the approach, a human instance annotation strategy is used to retrieve human segmentation annotations from multi-view RGB+D data and 3D human skeletons. In the experimental validation, such dataset has been used to fine-tune the state-of-the-art AIParsing network, by leveraging its instance-level annotations and multi-view data. The final model has been then evaluated on a subset of images from CIHP dataset with consistent overlaps between people, showing the effectiveness of the proposed approach, with an improvement in terms of body part-aware mean Intersection-over-Union up to 4.25% with respect to the original AIParsing network.

List of Figures

Contents

Chapter 1

Introduction

In recent years, human perception has gained relevant attention due to its fundamental role in many real-life applications, such as human-robot collaboration, virtual reality, video surveillance, social media and fashion editing. Multi-human parsing is a main computer vision task, allowing for rich human analysis in the wild.

1.1 Multi-human parsing

Human parsing, or human body parts segmentation, can be defined as the task of partitioning humans in an image into their semantic body parts (e.g., head, torso, arms and legs). Semantic classes of interest can sometimes also include clothing (e.g., shirt, pants, dress) and accessories. While human parsing approaches focus on segmenting body parts of a single person, in many real-world scenarios, the presence of multiple people is far more common. This kind of methods, however, can neither count the number of people in the image, nor can they determine to which person in the image each detected part belongs to. For this reason, recent efforts have been devoted to the development of techniques that aim at combining body part segmentation and human instance discrimination. This is known as multi-human parsing, as illustrated in Figure 1.1.

Multi-Human Parsing (MHP), or instance-level human parsing, aims at segmenting human body parts while simultaneously associating each part to the human instance (i.e., human body) such part belongs to. Multi-human parsing plays a critical role in many human-related tasks, in the fields of social media [1], fashion [2], human-robot collaboration [3] and human-centric vision [4, 5]. Several works, such as [6, 7], exploit human parsing to accurately detect clothing

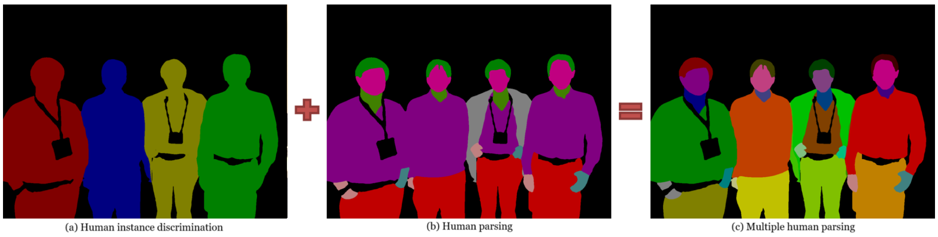


Figure 1.1: Multi-human parsing (c) can be considered as the combination of human instance discrimination (a) and human parsing (b) tasks.

and accessories, in order to develop fashion image applications, such as virtual clothes try-on. In the field of human-robot collaboration, body parts segmentation can be exploited to obtain a rich 3D semantic representation of people in the scene, useful to implement robust human collision avoidance, as described in [3]. Human parsing can also provide very useful cues and features for improved human pose estimation [8, 9], human action recognition, as in [5], and person re-identification [10, 11].

1.1.1 Challenges

Despite current multi-human parsing methods manage to achieve prominent results, benefiting from deep-learning advances, this task still poses many challenges. Since multi-human parsing is of interest for a very diverse range of downstream applications, in recent years, many models and datasets have been introduced. However, the vast majority of them is highly specific for the target domain. One of the main challenges of multi-human parsing is, indeed, represented by the lack of a common set of semantic classes of interest, resulting in a significant discrepancy in label granularity among different datasets. For example, the PASCAL-Person-Part (PPP) dataset [12] for instance-level human parsing is annotated considering 6 semantic classes (e.g., head, torso, upper/lower arms and upper/lower legs). This kind of annotations is typically used for applications which do not require prior knowledge about clothes or accessories, such as human-robot collaboration. Nevertheless, models designed for fashion purposes, for example, are interested in much more fine-grained categories. Crowd Instance-level Human Parsing (CIHP) [13] provides ground-truth pixel-level supervision for 19 different semantic classes, including 7 categories for body parts (i.e., hair, face, torso-skin, left/right arm, left/right leg) and 12 categories for

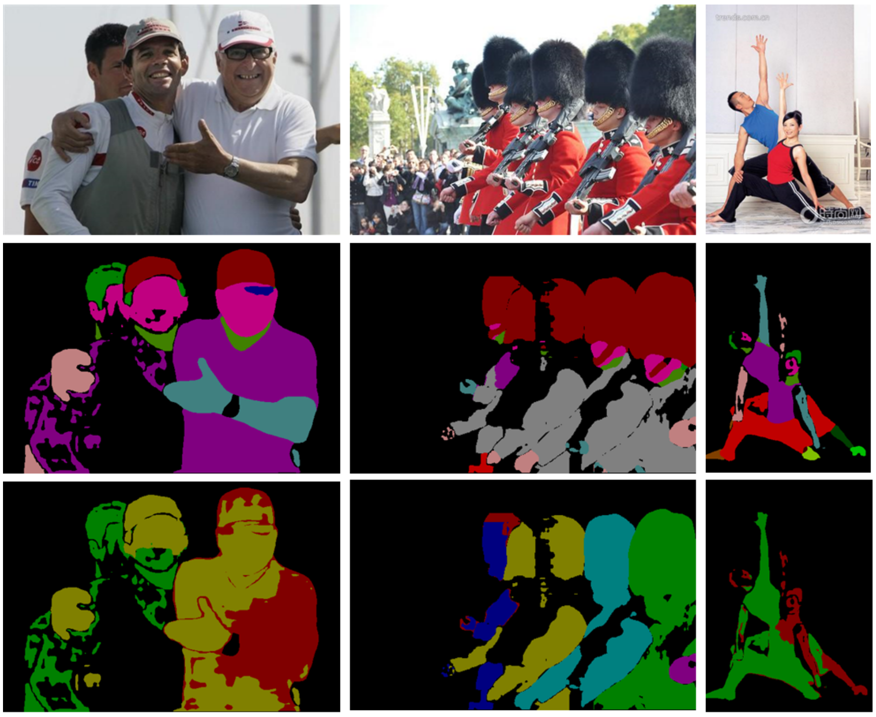


Figure 1.2: Example of multi-human parsing on images with overlapping human instances. RGB image (first row). Human parsing (second row). Human instance segmentation (third row).

clothes and accessories (i.e. hat, glove, sunglasses, upper clothes, etc.). Learning Vision Multi-Human Parsing (LV-MHP-v2.0) [14] is annotated with an even larger number of different parts, considering more than 50 different labels. Since classes of interest are not homogeneous between datasets, heavily restricts generalization of multi-human parsing models to different scenarios.

Another important issue is represented by intra-class variance, especially when considered classes include clothes and accessories. Objects belonging to the same semantic category (e.g. upper clothes) typically have very different appearances, with respect to color, texture and shape. Changes in illumination, viewpoint and resolution make correct classification even more challenging. On top of this, typical issues related to human perception, such as human body self-occlusions, unconstrained people pose and appearance, must all be taken into account. Having to deal with a varying number of people in each image, moreover, makes things even more complicated.

One of the most significant challenges of multi-human parsing, however, is represented by occlusions. Object occlusions have the effect of truncating human figures, resulting in incomplete human body structure. Occlusions between humans are far more challenging to solve, as shown in Figure 1.2. When occlusions create huge discontinuities in human body structure, current MHP methods fail to accurately detect human body parts, often producing very incomplete segmentation masks. In presence of multiple overlapping people, moreover, discriminating between different human instances represents a huge issue, as it becomes very difficult to understand to which person a given detected part belongs to. As illustrated in Figure 1.2, produced parsing maps are full of holes, many portions of the body are completely missed and several body parts are matched to the wrong person.

1.2 Contribution

As discussed in the previous subsection, despite recent progress and efforts in the field of multi-human parsing, several challenges still remain unresolved. In particular, this thesis focuses on addressing instance-level human parsing challenges in scenarios in which people are strongly overlapping. An in-depth analysis on current MHP methods highlights how heavy occlusions between people lead to severe deterioration in instance-level human parsing performance. In particular, when human bodies are largely overlapping, significant body portions are not segmented, leading to incomplete masks, and many body parts are matched to the wrong person, as it is difficult to distinguish between different human instances.

This thesis proposes to mitigate such issues exploiting information provided by multi-view systems. Multiple points of view on the same scene, in fact, provide different views on the same entity. This redundancy can be leveraged to retrieve the information lost because of occlusions. However, the lack of a multi-view dataset framing overlapping human instances with suitable instance-level body parts annotations, creates the necessity of producing an appropriate dataset to enable the development of the new multi-view approach. In order to do this, this work proposes an annotation strategy to produce accurate instance-level human segmentation masks from multi-view RGB+D video sequences, annotated with 3D body skeletons. The annotations produced do not include body part semantic information as very time-consuming to obtain. Furthermore, body part annotations are usually hand-made and thus, not very accurate. Considering the focus on overlapping people, in particular, this procedure is applied on the CMU Panoptic Studio dataset [15], providing images rich of overlapping people.

To exploit the available multi-view instance-level annotations, which are coarser with respect to typical supervision used for the MHP task, this thesis introduces a novel learning framework taking advantage of instance-level human segmentation as auxiliary task to enhance multi-human parsing in presence of strong occlusions between people in the image. Specifically, considering a multi-human parsing network and a target MHP dataset, both 2D and 3D human instance information from the novel dataset are exploited to enhance human instance segmentation and outline, as well as instance identity discrimination. In particular, a set of single-view human segmentation loss functions exploit 2D instance information to improve human segmentation quality in case of occlusions. A multi-view loss function improves human instance disambiguation between largely overlapped people in the image, exploiting 3D instance-level ground-truth to explicitly enforce human instance consistency between multiple views.

1.3 Thesis outline

The remainder of this thesis is organized as follows. In Chapter 2, the state-of-the-art about multi-human parsing is revised and analyzed, highlighting the main approaches adopted. In order to confirm and analyse the effects that scenarios presenting human overlaps have on multi-human parsing accuracy, in Chapter 3 an in-depth study on different multi-human parsing architectures is conducted. In Chapter 4 an annotation procedure to retrieve human instance-level segmentation annotations from RGB+D data is presented. Chapter 5 outlines the novel learning framework introduced, proposing to exploit human instance segmentation as auxiliary task to enhance multi-human parsing in presence of strong occlusions between people. Chapter 6 and 7 present the experimental results obtained validating the proposed approach. Chapter 8 concludes this thesis, summarizing contribution and illustrating future research directions.

Chapter 2

Related works

Multi-Human Parsing (MHP) finds many applications in the fields of multimedia and computer vision, enabling fine-grained, pixel-wise human analysis. It is of strategic importance in many human-related tasks, such as human action recognition [5], human-robot collaboration [3], human pose estimation [4,8,16], person re-identification [17,18] and fashion image manipulation [2,6,7].

Multi-human parsing can be considered as the combination of two different computer vision tasks: human instance segmentation and human parsing. The former consists in distinguishing and localizing human bodies in an image, while the latter partitions each human body into accurate semantic body parts, such as head, torso, arms and legs. Semantic classes of interest sometimes include also clothing and accessories.

This chapter is organized as follows. The first part will introduce the tasks of instance segmentation and human parsing, to give the reader a better understanding of the task at hand. In the second part of the chapter, the state-of-the-art of multi-human parsing will be revised and analysed, presenting different deep-learning-based approaches proposed in the literature. Finally, the most popular multi-human parsing datasets will be introduced and described.

2.1 Instance segmentation

Instance segmentation is the task of identifying all occurrences of a certain object of interest in an image, while producing an accurate segmentation mask for each different instance of that object [19]. It therefore combines two different perception tasks, namely *object detection*, that classifies and localizes objects in the image, and *semantic segmentation*, that predicts a semantic category for each pixel. For instance segmentation, each object region extracted by the detection are given as input to semantic segmentation, that produces accurate masks for each detected object. Current instance segmentation approaches are entirely based on deep-learning techniques. Proposed methods can be divided into two main families: region proposal-based methods and fully convolutional methods. Most current multi-human parsing techniques build on instance segmentation architectures belonging to one of these two categories.

2.1.1 Region proposal-based instance segmentation

Region proposal-based instance segmentation approaches produce segmentation masks for candidate object regions extracted from the image. These techniques usually build on existing object detectors, such as Faster R-CNN [20]. A prominent example of this class of approaches is represented by Mask R-CNN [19], a pioneer work proposed in 2017. Despite not being the current state-of-the-art, Mask R-CNN is still very popular as a strong and reliable starting point for instance segmentation applications, such as multi-human parsing. Mask R-CNN extends a well-known object detection framework, Faster R-CNN [20], adding a new branch based on Fully Convolutional Networks (FCNs) [21] to predict accurate segmentation masks for each detection proposal. The overall architecture of Mask R-CNN for instance segmentation is depicted in Figure 2.1. The Faster R-CNN detector [20] works in two stages. The first stage is called Region Proposal Network (RPN) and predicts bounding boxes for candidate objects. These are known as Regions of Interest (RoIs). The second stage, coming from precursor Fast R-CNN [22], extracts features from objects proposals to perform box classification and regression. Adding to this, Mask R-CNN augments the second stage with a novel branch, producing binary segmentation masks, one for each object category, for each region. To adapt the object detection framework to semantic segmentation, the Mask R-CNN architecture proposes *RoIAlign*, a simple layer used to preserve and retrieve the spatial location of features extracted from object proposals. This is essential to pixel-to-pixel alignment between network inputs and the outputs.

For good quality features extraction, many instance segmentation techniques exploit a combination of convolutional architectures, such as ResNet [23], and Feature Pyramid Networks (FPNs) [24], shown in Figure 2.2. A feature pyramind network essentially builds a pyramid of feature levels from a single-scale input. This allows to extract features for a given RoI from different levels of the pyramid,

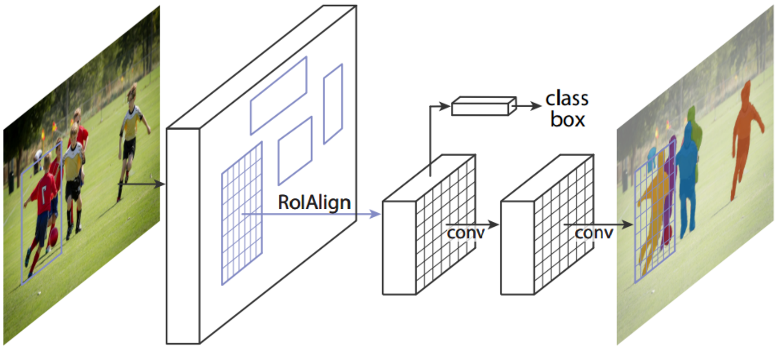


Figure 2.1: Overview of Mask R-CNN framework for instance segmentation [19].

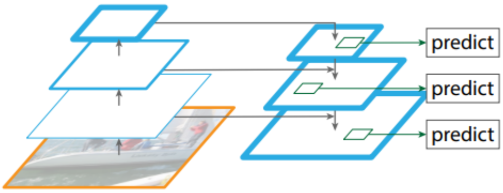


Figure 2.2: Overview of Feature Pyramid Network (FPN) architecture, for multi-scale feature maps generation [24].

according to region scale. This leads to high quality feature maps, of utmost importance for accurate bounding boxes and masks generation.

An evolution of Mask R-CNN is represented by Path Aggregation Network (PANet) [25], shown in Figure 2.3. While lower levels in feature pyramid networks are very useful to identify and describe large scale object instances, the path from such feature levels to top features is rather long, resulting in the loss of accurate information. Furthermore, the assignment of candidate regions having different scales to corresponding feature pyramid levels, follows heuristics. To mitigate these issues, PANet enhances the feature pyramid with accurate localization information, propagating features coming from lower, higher resolution layers up

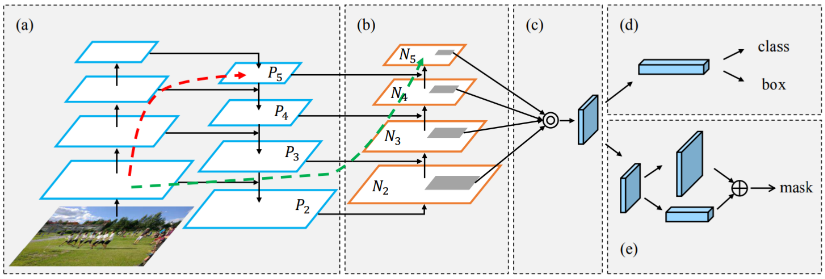


Figure 2.3: Overview of Path Aggregation Network (PANet) architecture for instance segmentation [25]. (a) Feature Pyramid Network (FPN) backbone. (b) Bottom-up path augmentation. (c) Adaptive feature pooling. (d) Box prediction branch. (e) Fully-connected fusion.

to the topmost feature levels. To avoid arbitrarily assigning regions of interest to feature scales, an adaptive feature pooling strategy is introduced, consisting in aggregating features coming from all pyramid levels to produce a rich feature map for each region proposal, as shown in Figure 2.3.

2.1.2 Fully-convolutional instance segmentation

In contrast with the instance segmentation approaches presented so far, these methods do not rely on a region proposal stage, but make use of fully convolutional architectures. These techniques do not build on object detectors, like Mask R-CNN [19] or PANet [25], but perform instance segmentation in a single stage. For this reason, they typically achieve high inference speed, independently from the number of people in the image, which makes such approaches very interesting to multi-human parsing methods.

A prominent example of fully convolutional solutions for the instance segmentation task is represented by YOLACT, You Only Look At CoefficienTs [26]. YOLACT is able to achieve real-time inference by decomposing the instance segmentation problem into two parallel subtasks: the generation of a dictionary of prototype masks over the entire image and the prediction of mask coefficients for each instance, to produce a linear combination of the generated prototypes. The overall architecture of YOLACT is shown in Figure 2.4.

Other fully-convolutional instance segmentation methods are based on fully-convolutional object detection, provided by frameworks such as FCOS [27]. FCOS (Fully-Convolutional One-Stage) is a popular anchor-free object detection framework. Anchor-based detection, despite being very widely adopted [28, 29], has several drawbacks. This kind of techniques generate a series of box candidates depending on predefined anchor box sizes, aspect ratios and scales. All these

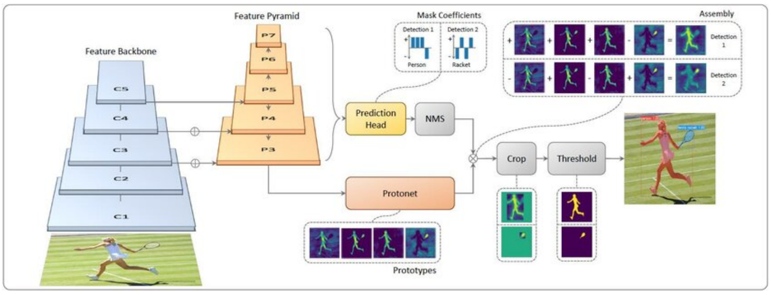


Figure 2.4: Overview of the architecture of YOLACT instance segmentation network [26]. Blue regions in the prototype maps indicate low confidence values, while yellow regions indicate high confidence values.

hyperparameters, however, need accurate tuning, which restrains the generalization of such models to new datasets and detection tasks. Moreover, anchor-based object detectors typically exploit architectures that are not based on pixel-level predictions, which is inconsistent with what is required by segmentation tasks. In contrast, anchor-free frameworks, such as FCOS, do not require many hyperparameters and solve object detection in a per-pixel prediction fashion, which makes them very suitable to be used as starting point for the development of instance segmentation applications. Given this, FCOS-based instance segmentation is recently being exploited by successful multi-human parsing approaches, such as [30].

2.2 Human Parsing

Human Parsing, or body parts semantic segmentation, is a fine-grained segmentation task which consists in partitioning a human body image into its semantic body parts, such as head, torso, arms and legs [31]. With respect to typical semantic segmentation tasks, human parsing presents several additional challenges. In many applications, predictions are required to be extremely fine-grained, differentiating between left and right parts of the body and distinguishing between several types of garments and accessories. In addition to this, most of the categories considered by this task suffer from a rather large intra-class variance. A semantic class such as ‘upper clothes’, for instance, aggregates pieces of clothing that can be very different in color, texture and shape. Further challenges are represented by variations in illumination and viewpoint, low resolution images and unconstrained human poses, that can easily compromise the accurate distinction between right and left body parts.

Human parsing models can be distinguished considering how they model the relationship between human body parts, in order to extract meaningful features. Graph Pyramid Mutual Learning (Grapy-ML), introduced by He et al. [32], for instance, propose to take advantage of the attention mechanism. Self-attention is exploited to model the connections between body part nodes in order to address human parsing at different levels of granularity, as shown in Figure 2.5.

Other approaches, such as [31] take advantage of auxiliary tasks in order to better model the relationship among body parts. Edge information, for example, provides useful cues about body parts boundaries and significantly enhances the ability of the model in discriminating between adjacent parts. Following

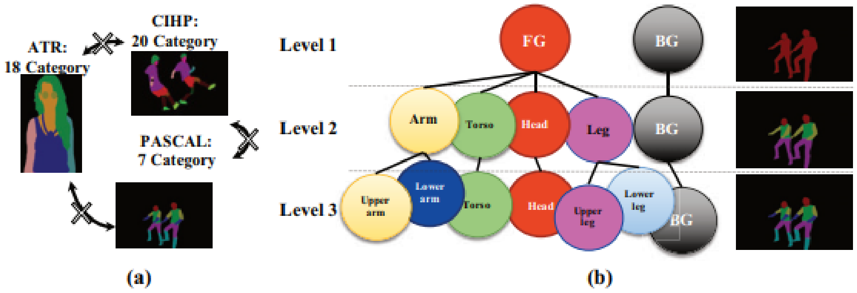


Figure 2.5: Multi-granularity lexical pyramid representation of the human body considered by Grapy-ML human parsing model [32].

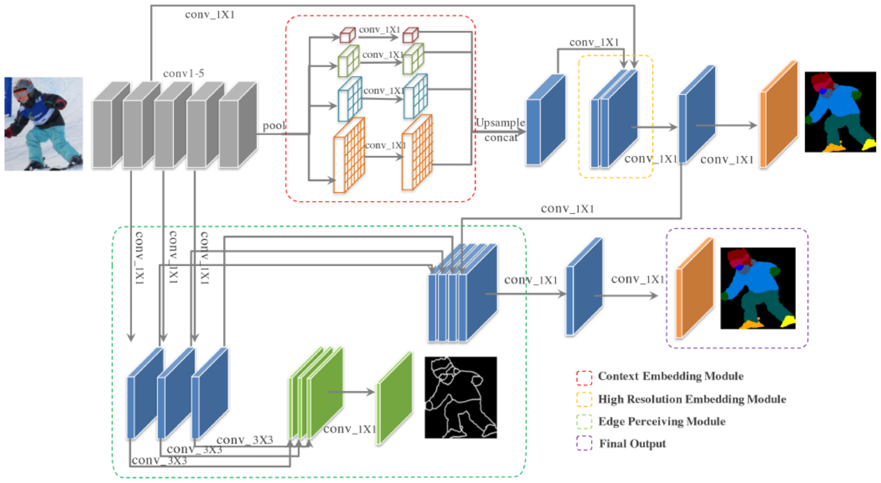


Figure 2.6: Overall architecture of the Context Embedding with Edge Perceiving framework for human parsing [31]. The framework consists of three key modules: 1) high resolution embedding module 2) global context embedding module, 3) edge perceiving module.

this intuition, Ruan et al. proposes a Context Embedding with Edge Perceiving framework, known as CE2P [31], shown in Figure 2.6. The network architecture essentially consists of three key modules. At first, a high-resolution context embedding module enlarges the feature map and preserves fine-grained details about the human body that needs to be segmented. Then, a global context embedding module encodes global information via multi-scale features. Finally, an edge perceiving module takes into account the contours of the body parts to produce accurate results.

Given the level of granularity expected from human parsing predictions, as mentioned at the beginning of this subsection, annotations for this task are required to be very precise, making ground-truth generation a rather expensive process. Moreover, most current approaches are entirely based on deep learning techniques, which notoriously require a lot of annotations for effective supervised

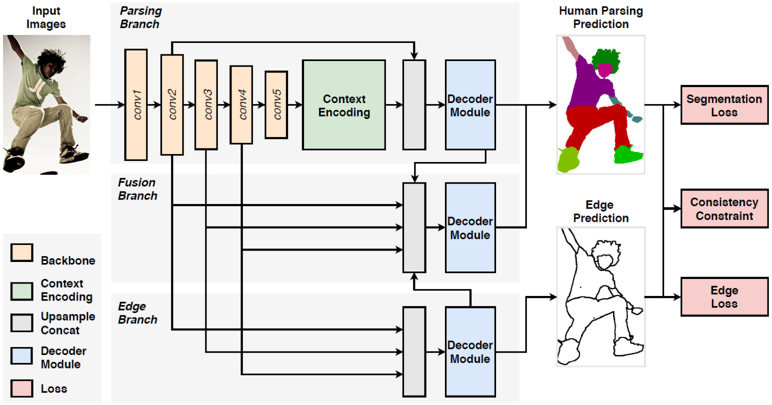


Figure 2.7: Overview of the Self-Correction for Human Parsing (SCHP) framework [33].

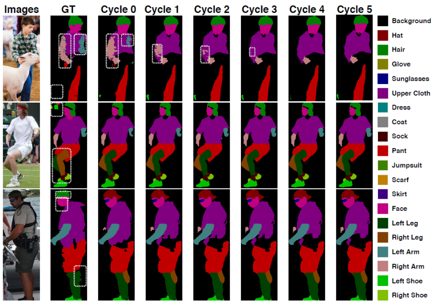


Figure 2.8: Examples of SCHP self-correction mechanism on ground truth annotations [33].

training. For these reasons, human parsing architectures often have to deal with noisy ground truth annotations, which can be detrimental to overall model performance. Trying to solve this issue, Li et al. build upon CE2P [31] and propose a purification strategy named Self Correction for Human Parsing (SCHP) [33], shown in Figure 2.7, which progressively refines ground-truth labels, making them more reliable for supervision as training goes on. At each training iteration, the current estimated model is aggregated with the former optimal one and used to produce more reliable masks, as depicted in Figure 2.8 for few sample images.

2.3 Instance-level Human Parsing

While human parsing methods focus on segmenting human body parts of a single human instance, in many real-world scenarios, the presence of multiple people in an image is far more common. Instance-level human parsing techniques aim at segmenting human body parts while also associating each part to the human instance it belongs to. Multi-human parsing is indeed a difficult task, combining challenges typical of single human parsing, such as self-occlusions, variance in people appearance and pose, with struggles brought by the fact of having to deal with a varying number of people, potentially occluding and overlapping with one another. Multi-human parsing methods proposed in the literature can be distinguished into two main classes: bottom-up and top-down approaches. Current methods are based on deep-learning techniques, requiring huge datasets to be trained. In the following subsection, the most popular datasets will be introduced, while the remainder of this section will be devoted to presenting the main works on bottom-up and top-down multi-human parsing.

2.3.1 Datasets

Instance-level human parsing models are typically trained and evaluated on two popular datasets, Crowd Instance-level Human Parsing (CIHP) [13] and Learning Vision Multi-Human Parsing (LV-MHP-v2.0) [14].

CIHP Crowd Instance-level Human Parsing (CIHP) [13] is the largest publicly available dataset for multi-human parsing in the wild. It includes a total of 38,280 images, annotated considering instance-level identification and 19 human semantic parts, including body parts (e.g., face, torso, left/right arm, etc.), but also clothing and accessories (e.g., coat, dress, hat, glove, etc.). Images have been collected from sources such as Google and Bing research engines, and frame people appearing in different poses, viewpoints and in a wide range of resolutions. The dataset is split into three parts. The training set includes 28,280 images, while the validation and the test set contain 5,000 images each. Figure 2.9 shows some examples of images and annotations extracted from the CIHP dataset, while Figure 2.10 illustrates dataset statistics about the number of people appearing in the images and about the data distribution of the 19 annotated semantic parts labels.

LV-MHP-v2.0 Learning Vision Multi-Human Parsing (LV-MHP-v2.0) [14] contains 25,403 images, annotated considering 58 fine-grained semantic categories, including body parts, clothing and accessories. The training set includes 15,403 images, while two splits of 5,000 images each are reserved for validation and testing, respectively. The dataset frames a minimum of 2, up to a maximum of 26, people per image and captures real-world scenes from various viewpoints, with a wide range of poses, interactions and backgrounds. Figure 2.11 illustrates some examples of images and ground-truth masks extracted from the dataset.

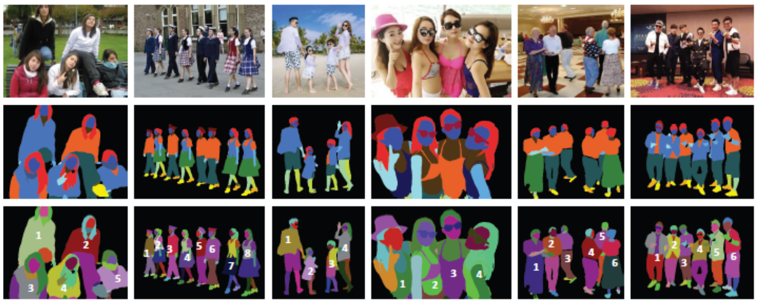


Figure 2.9: Examples from the large-scale "Crowd Instance-level Human Parsing (CIHP)" Multi-Human Parsing dataset. Images are presented in the first row. Semantic part segmentation and instance-level human parsing are shown in the second and third row respectively [13].

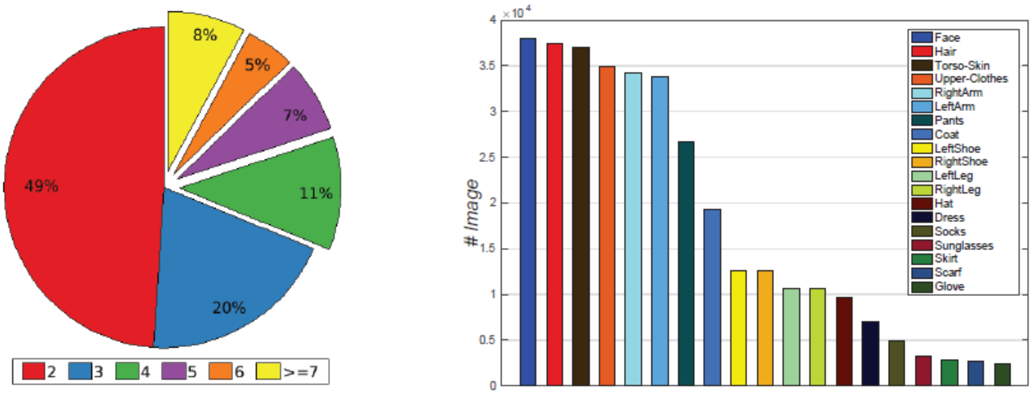


Figure 2.10: On the left, statistics on the number of persons in one image and on the right the data distribution on the 19 semantic part labels in the CIHP dataset [13].

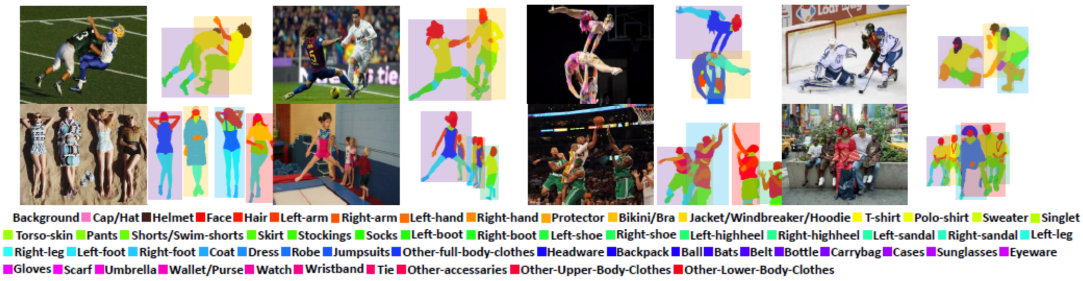


Figure 2.11: Examples from the large-scale "Learning Vision Multi-Human Parsing (LV-MHP-v2.0)" Multi-Human Parsing dataset [14].

2.3.2 Bottom-up approaches

Bottom-up approaches primarily consider multi-human parsing as a fine-grained semantic segmentation task, whose objective is to predict pixel-level body part categories. Subsequently, classified pixels are grouped into different human instances. This can be done following different strategies, such as edge-aware clustering, as in [13] or mapping body parts to body joints, as in [8]. These type of methods usually show better performances in terms of body parts segmentation, but have the tendency to confuse adjacent and overlapping human instances.

Part Grouping Network (PGN) [13], shown in Figure 2.12, is the first work that defines instance-level human parsing as the composition of two twinned subtasks: part-level pixel grouping and instance-level part grouping. Part-level pixel grouping consists in assigning to each pixel in the image the corresponding body part label, achieved using human parsing techniques. Instance-level part grouping matches the segmented semantic parts to the human instances they belong to. In this case, this is done by exploiting human instance boundary cues given by human body edges detection. Body parts contained by human edges, indeed belong to that instance.

Zhou et al. in [8] propose a different part-grouping strategy, shown in Figure 2.13. The proposed network aims at solving the instance-level human parsing task at multiple levels of granularity, jointly learning multi-human pose estimation and multi-human body parts segmentation. Adopting a dense-to-sparse projection field, dense body parts predictions are mapped to sparse body keypoints, casting the part-grouping problem to a multi-person body joint composition task. Semantic parts are associated to the human body whose joints are the closest, effectively fusing knowledge about human pose and pixel-level part semantics.

Transformer-based approaches

Among bottom-up approaches, transformer-based methods exploit the latest advances in the deep-learning field to enhance multi-human parsing. Yang et al. in [34] proposes Mask2Former for Parsing (M2FP), a new transformer-based base-

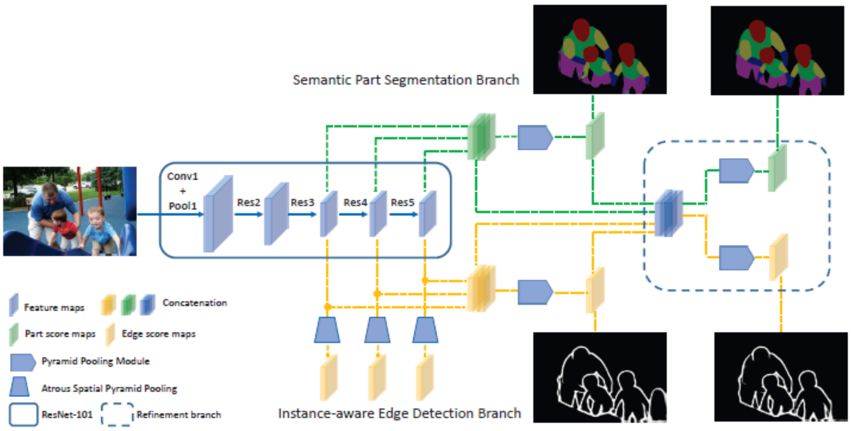


Figure 2.12: Overall architecture of Part Grouping Network (PGN) [13].

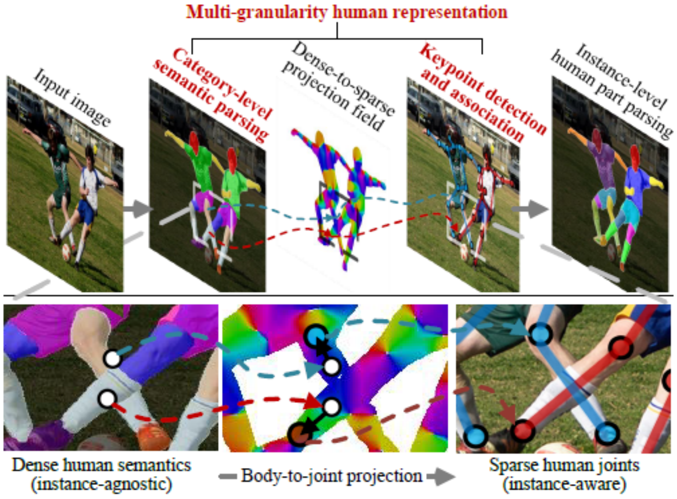

Multi-granularity human representation
Input image
Category-level semantic parsing
Dense-to-sparse projection field
Keypoint detection and association
Instance-level human part parsing
Dense human semantics (instance-agnostic)
Body-to-joint projection
Sparse human joints (instance-aware)

Figure 2.13: Overview of Differentiable Multi-Granularity Human Representation Learning for Instance-Aware Human Semantic Parsing [8].

line for multi-human parsing, built on the Mask2Former architecture [35], proposed for universal image segmentation. An overview of the architecture is illustrated in Figure 2.14. Mask2Former for Parsing takes advantage of the human body hierarchy and of the powerful sequence encoding capabilities of transformer-based methods to model hierarchical relationships between body parts and humans instances. In particular, it makes use of three different kinds of queries: background queries, part queries and human queries. The self-attention mechanism is, therefore, used to learn relationships between body parts, between human instances, between body parts and human instances and between parts, human instances and background. M2FP can be applied to both human parsing and multi-human parsing tasks, yielding very good performances in terms of both accuracy and inference speed.

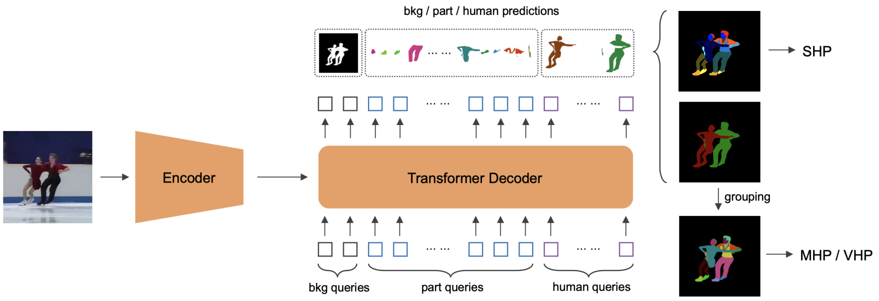


Figure 2.14: Overview of the architecture of the Mask2Former for Parsing (M2FP) network [34].

2.3.3 Top-down approaches

With respect to bottom-up methods, top-down multi-human parsing techniques first locate people in the image and then independently segment body parts for each human instance. In this category, the literature differentiates between two-stage top-down and one-stage top-down approaches.

Two-stage top-down approaches

Two-stage top-down approaches to multi-human parsing start from a well-trained object detector, typically Mask R-CNN [19], and apply a robust single human parsing method to the detected human bodies. This category of methods mainly focus on the body part segmentation logic and rely on state-of-the-art techniques for the detection stage. Approaches adopting this kind of architecture exhibit higher accuracy in the instance-level human parsing task, at the expense of inference speed and flexibility, which represent the bottleneck for this family of methods.

Ruan et al. in [31] show that the proposed single human parsing network CE2P, presented in Section 2.2, can be easily adapted to the multi-human scenario exploiting Mask R-CNN to locate different people in the image (M-CE2P). Then, instance-level body parts segmentation is obtained by combining the result of two parallel branches, devoted to *global* and *local* parsing, respectively. Global parsing consists in applying the CE2P network to the image as a whole, to retrieve global context information, and the output of this step is fused with the result from the local parsing branch, that applies the human parsing network to human patches extracted by the object detector.

In a similar fashion, different human parsing techniques such as [36, 37], as well as those presented in Section 2.2 can all be extended to the multi-human case combining the human parsing network with a well-trained object detector.

One-stage top-down approaches

Similarly to two-stage methods, one-stage top-down multi-human parsing architectures locate human instances first and then parse each body in a fine-grained manner. However, in contrast to the former approaches, for one-stage top-down methods detection and human parsing sub-networks are trained together. This kind of networks are very flexible and can be easily adapted to different downstream tasks by adding new modules. In terms of accuracy and inference speed

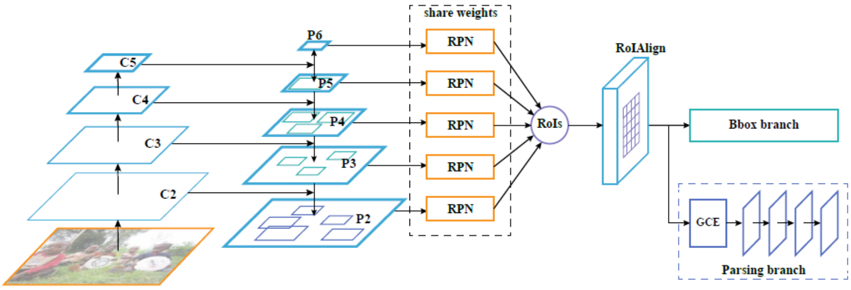


Figure 2.15: Overview of Parsing R-CNN network for multi-human parsing [16].

they represent a good compromise between bottom-up approaches and two-stage methods and, for this reason, they are currently the mainstream research direction for instance-level human parsing [38].

One of the first successful attempts at employing this kind of architecture comes from Yang et al. [16]. Parsing R-CNN aims at improving instance-level human analysis working on feature quality. An overview of the Parsing R-CNN architecture is shown in Figure 2.15. To enhance the semantic information provided to the network, Parsing R-CNN adopts a proposals separation sampling strategy. In FPN [24] and Mask R-CNN [19] large regions of interest will be assigned to coarser-resolution feature maps, while small regions will be assigned to feature maps having finer resolution. However, since human instances usually occupy a large portion of images, this solution is not optimal for instance-level human parsing, as most operations would be performed on feature maps that do not retain many instance details. To overcome this issue, Yang et al. propose in [16] to extract features used by the parsing branch from the finest-resolution feature map only, while keeping the standard region assignment rule for the bounding box detection branch. At the same time, to preserve as many details as possible, the architecture introduces a Geometric and Context Encoding module (GCE) to enlarge the receptive field and capture relationships between body parts, combining Atrous Spatial Pyramid Pooling (ASPP) [39] and non-local operators [40].

Yang et al. build upon this work and present Renovating Parsing R-CNN (RP R-CNN) [38], shown in Figure 2.16. In this work, feature semantics is boosted even further introducing a Global Semantic Enhanced Feature Pyramid Network (GSE-FPN). Starting from the popular feature pyramid network (FPN) [24], generated multi-scale features are up-sampled to the same scale and fused, strengthening and propagating the effect of global information on extracted features, benefiting human parsing accuracy. RP R-CNN also introduces a novel

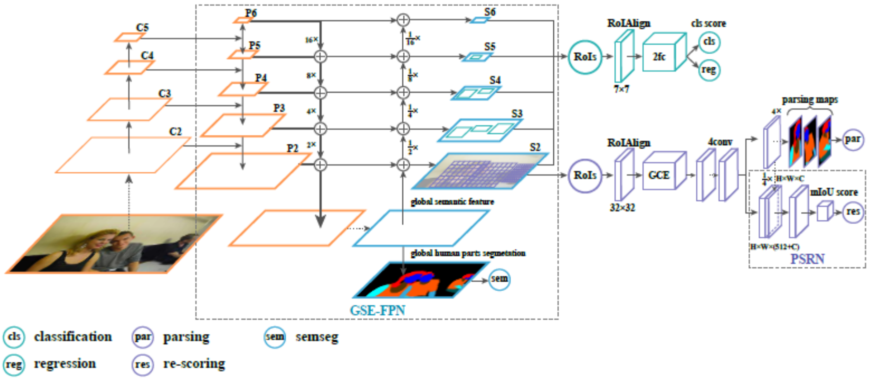


Figure 2.16: Overview of RP R-CNN network architecture [38].

parsing re-scoring network, in order to measure the quality of produced instance parsing maps, using the computed score to easily filter out poor results.

While the majority of available approaches employ anchor-based detectors for human bounding box regression, Zhang et al. recently proposed Anchor-free Instance-level Human Parsing (AIParsing) [30], which exploits the fully-convolutional anchor-free detection head FCOS [27], to localize human instances. The overall architecture of the network is shown in Figure 2.17. Given an input image, a combination of a CNN-based backbone (i.e., ResNet [23]) and a feature pyramid network [24] is used to extract multi-scale features. These are fed to the anchor-free detection head, which predicts a bounding box for each human instance. The final instance parsing maps are then the result of the application of an edge-guided parsing head to the detected boxes.

AIParsing is currently the state-of-the-art among one-stage top-down ap-

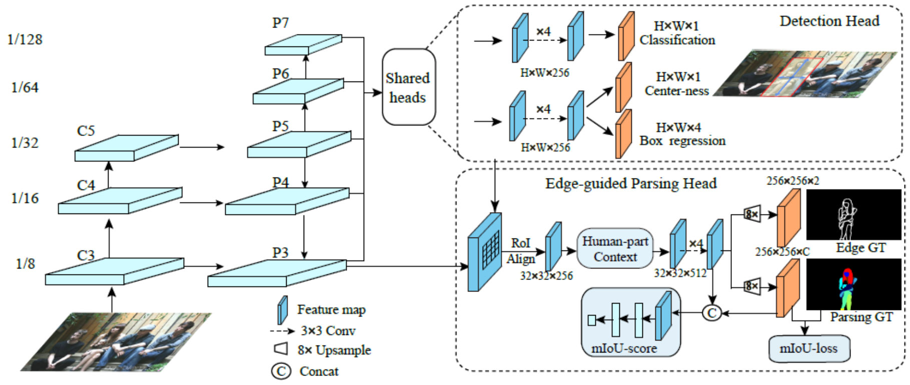


Figure 2.17: Overall architecture of Anchor-free Instance-level Human Parsing network [30].

proaches and has been considered as starting point for the development of the approach proposed in this thesis. This network will be now described more in detail, to give some concepts useful to illustrate the devised multi-human parsing method in Chapter 5. AIParsing is mainly composed of two sub-networks: an anchor-free human detection head and an edge-guided parsing head, with edge information used to distinguish adjacent human body parts. A refinement head is also added to enhance the final parsing results. The loss of the detection head is defined as follows:

$$L_{det} = L_{cls} + L_{reg} + L_{center} \tag{2.1}$$

where L_{cls} is the classification loss, L_{reg} is the box offset regression loss and L_{center} is the loss on box centerness. L_{cls} is the focal loss [41], L_{reg} is the bounding box Intersection over Union loss, as defined in [42] and L_{center} is the binary cross entropy loss. The edge-guided parsing head takes advantage of edges extracted from the image to distinguish adjacent body parts and human instances, particularly useful when dealing with overlapping human bodies. The edge-guided parsing head, shown in Figure 2.17, is the combination of three main parts. A detail-preserving component has the goal of extracting high quality features to retain important human appearance information, very useful for instance parsing. Features corresponding to detected instances, used to predict the final segmentation, are extracted from the lowest layer of the feature pyramid, to exploit high resolution information. On the outputs of this detail-preserving feature extraction step, the human-part context encoding is used to capture information about context, of foremost importance for semantic segmentation tasks. This is achieved by employing the Pyramidical Gather-Excite Context module (PGEC) [36], extracting multi-scale information, in combination with non-local operators [40], to capture spatial relations and provide information about the relative position between different parts of the body. The edge-guided parsing head is trained with the following multi-task loss function:

$$L_{pred} = \alpha L_{parsing} + \beta L_{edge} \tag{2.2}$$

where α and β are both set to 2. $L_{parsing}$ is the standard cross-entropy loss on body parts segmentation. L_{edge} looks at body parts boundaries and it is the weighted cross-entropy loss, defined as follows:

$$L_{edge} = -\omega_0 \sum_{i \in Y_-} log((p_i(y_i = 0)) - \omega_1 \sum_{i \in Y_+} log((p_i(y_i = 1)) \tag{2.3}$$

where Y_+ and Y_- are the ground-truth pixels belonging to edges or non-edges, ω_0 is equal to $\frac{|Y_+|}{|Y|}$, ω_1 is $\frac{|Y_-|}{|Y|}$ and p_i indicates the probability of the i-th pixel. The refinement head is used to improve both human instance and parsing map accuracy. To discard imprecise human detection results, a re-scoring sub-network computes the mean Intersection over Union of the predicted parsing map within the detected bounding box, assigning a score to each detection. Poor scoring boxes are discarded. Simultaneously, the quality of parsing maps is optimized, exploiting the Lovász-Softmax loss, directly optimizing the Intersection over Union measure [43]. The loss of the refinement head is:

$$L_{refine} = \theta L_{miou} + \gamma L_{miou-score} \tag{2.4}$$

where γ is equal to 1 and θ is equal to 2. L_{miou} is the mIoU loss and $L_{miou-score}$ is the mIoU score for box quality. The total loss L_{total} used to train AIParsing is the sum of the detection head loss L_{det}, the prediction head loss L_{pred} and the refinement head loss L_{refine}. The network has been evaluated on two popular multi-human parsing datasets, CIHP [13] and LV-MHP-v2.0 [14]. The network is trained using Stochastic Gradient Descent (SGD) for about 75 epochs with batches of 8 images. The initial learning rate is set to 0.005 and decreased by a factor of 10 after 50 and after 65 epochs. The weight decay is set to 0.0001 and the momentum is equal to 0.9. ResNet-101 is used as backbone for feature extraction, initialized with ImageNet pre-trained weights [44]. On both datasets, AIParsing achieves state-of-the-art performances, topping other one-stage top-down alternatives.

Chapter 3

Analysis of multi-human parsing methods on overlapping human instances

One of the main challenges that instance-level human parsing methods need to tackle is represented by overlapping bodies. Despite being generally good at segmenting people occluded by objects [34], current MHP techniques struggle to correctly detect and segment human instances when they overlap with each other. As described in Section 2.3, there exist two main types of multi-human parsing architectures, bottom-up and top-down approaches, both affected by this issue. Bottom-up approaches consider MHP primarily as a body parts segmentation task, and then associate the segmented parts to the human instance they belong to. This last stage is the most problematic when in presence of overlapping people, as occlusions cause ambiguity and the model struggles in matching each part with the correct person. Differently, top-down approaches first detect human instances, regressing a bounding box for each different human body, and then parse extracted regions. As a consequence, in case of occlusions, both detection and parsing are affected. During the detection stage, the model might fail in detecting bodies which are largely occluded, leading to missed instances in the final prediction. During the parsing stage, the problem is different. The model focuses on single extracted boxes that, when people are overlapping, will contain body parts which do not belong to the target instance. The issue, in this case, becomes identifying the main person in the region and considering the others as part of the background.

Given these premises, this chapter presents a study on the effect of overlapping human instances on the performance of three different multi-human parsing architectures, introduced in Section 2.3: two top-down approaches, namely RP R-CNN [38] and AIParsing [30], and bottom-up approach Mask2Former for Parsing (M2FP). Considering task-specific evaluation metrics, described in the following section, these techniques will be evaluated under different occlusions scenarios, to investigate how the amount of overlap between people in the image impacts on the performances.

3.1 Evaluation metrics for instance-level human parsing

To evaluate multi-human parsing methods, two types of metrics are usually considered: *global-level metrics* and *instance-level metrics* [30]. Global-level metrics measure the quality of body parts segmentation. These are consistent with those typically used for the semantic segmentation task. They include pixel accuracy, mean pixel accuracy and mean Intersection-over-Union. Pixel accuracy (pix_acc) is the ratio between the number of pixels whose category was correctly predicted and the total number of pixels in the image. In other words, it represents the percentage of adequately classified pixels. Even though this seems a reasonable evaluation metric for semantic segmentation, it can sometimes provide misleading results, especially if the classes of interest take up a small portion of the image with respect to background, as often happens for human parsing and multi-human parsing tasks. A simple improvement on this metric is represented by mean pixel accuracy (mean_acc), that computes pixel accuracy for each category and then considers the average of the obtained values. The most relevant among global-level evaluation metrics is, however, mean Intersection-over-Union (mIoU). Intersection-over-Union, or Jaccard index, quantifies the percentage of overlap between the ground-truth segmentation mask and the predicted output. It calculates the intersection and the union of two sets: the ground-truth and the predicted segmentation, and then considers the ratio between these two quantities. Mean IoU simply computes the Intersection-over-Union for each class and then takes the mean of such values.

Instance-level evaluation metrics measure the performance of actual instance-level human parsing. Average Precision based on Part (AP^P) evaluates part segmentation considering also human instances. In particular, for each instance, a prediction is regarded as correct (i.e., a true positive) if the mean Intersection-over-Union between ground-truth masks and predicted parts is higher than a certain threshold. At this point, the Average Precision is given by the area under the Precision-Recall curve, where Precision is the ratio between correct and total predictions, whereas Recall represents the ratio between correct predictions and ground-truth. AP^P_{50} considers the threshold for Intersection-over-Union to be 0.5, AP^P_{vol} is the mean of the results obtained using various thresholds, usually from 0.1 to 0.9 with a 0.1 step. Probability of Correct Parts (PCP) gives a measure of the human parsing quality within the human instance. For each true positive human body, it considers as correct predictions those semantic categories, excluding background, having IoU greater than a threshold, typically 0.5. PCP for the given instance is the ratio between the number of correctly parsed semantic categories and the total number of categories of that same person, according to ground-truth.

3.2 Impact of overlapping human instances on multi-human parsing performance

To investigate the impact of overlapping human instances on instance-level human parsing, various models will be tested on popular MHP datasets, namely Crowd Instance-level Human Parsing (CIHP) [13] and Learning Vision Multi-Human Parsing (LV-MHP-v2.0) [14]. Specifically, different subsets including images presenting different degrees of overlap severity will be considered. For each dataset, the validation split is divided into four sets, depending on the degree of overlap between people in the images. The degree of overlap, referred to as DoO, for a given image is defined as the Intersection-over-Union between human bodies, computed considering ground-truth bounding boxes. The subsets considered for the analysis are as follows:

- images with degree of overlap greater or equal than 0.2 (DoO_20);
- images with degree of overlap greater or equal than 0.4 (DoO_40);
- images with degree of overlap greater or equal than 0.6 (DoO_60);
- images with degree of overlap greater or equal than 0.8 (DoO_80).

RP R-CNN [38] and AIParsing [30] are used to represent the class of top-down approaches, with the former relying on classic anchor-based detection techniques

	N. Images
Validation set	5000
DoO_20	2550
DoO_40	1158
DoO_60	318
DoO_80	55

Table 3.1: Composition of subsets from CIHP validation set for different values of degree of overlaps.

and the latter being based on anchor-free detector FCOS [27], as described in Section 2.3. For bottom-up approaches, Mask2Former for Parsing (M2FP) [34] is selected. Each model is tested on the identified subsets of images and performance is evaluated considering mean Intersection over Union as global-level metric and Average Precision based on Parts (AP^P_{vol}) as instance-level metric.

3.2.1 Analysis on CIHP dataset

For CIHP, the subsets considered are as illustrated in Table 3.1. Evaluation results for RP R-CNN, AIParsing and M2FP are shown in Table 3.3. Considering the variation of the mean Intersection-over-Union over the different sets, it can be observed that the quality of human parsing monotonically decreases as the degree of overlap becomes larger. This means that the more people in the image occlude each other, the more the models struggle to accurately segment human body parts. In particular, it is interesting to notice that state-of-the-art AIParsing performs strictly worse than RP R-CNN, for which the degradation in human parsing quality is slower. Mask2Former for Parsing performs a lot better than the top-down approaches on the whole validation set and, reasonably, the advantage is preserved even in case of occlusions between people in the image. However, even in this case, occlusions are detrimental for the model performance, as in the case of images with 80% of instance overlap, where performance drops by 16.1%.

Moving to instance-level performance evaluation, the trend in the results is similar, showing that instance-level human parsing quality decreases as overlaps become more severe. Comparing top-down approaches, the slope with which performances go down is very similar, therefore, in this case, AIParsing, that surpasses RP R-CNN on the whole validation set, works slightly better when encountering overlaps. Looking at the bottom-up approach M2FP, the decrease in instance-level segmentation quality is present, but is not as prominent, proving that for this class of methods, overlaps still represent a challenge but not as critical as it is for top-down techniques.

	N. Images
Validation set	5000
DoO_20	2353
DoO_40	1306
DoO_60	530
DoO_80	165

Table 3.2: Composition of subsets from LV-MHP-v2.0 validation set for different values of degree of overlaps.

3.2.2 Analysis on LV-MHP-v2.0 dataset

For the LV-MHP-v2.0 dataset, the composition of the subsets considered is described in Table 3.2. Evaluation results for the different models are shown in Table 3.5.

As this dataset is extremely challenging, presenting annotations for more than 50 different semantic body parts, global human parsing is generally very difficult. For this reason, examining how the mean Intersection-over-Union varies according to the degree of overlap, the detrimental effect given by occlusions is less evident with respect to the previous dataset. However, considering instance-level metrics, the negative impact of the overlaps becomes very visible. In particular, consistently to what can be observed for CIHP, top-down methods struggle more and more as occlusions get more severe, while the impact for the bottom-up approach is less significant. When considering images having DoO equal to 80, for instance, Average Precision for AIParsing drops by 18.8%, while only by 7.1% for M2FP.

Analyses conducted on CIHP and LV-MHP-v2.0 highlight that strong overlaps between human instances represent a huge problem for current multi-human parsing approaches. How to effectively deal with occlusions between people appears to be still very unresolved and should be urgently addressed. In particular, comparing results for the evaluated models, it is clear that this challenge prominently affects top-down architectures, bringing destructive effects on performances. Motivated by this, in the remainder of this thesis, a novel approach for multi-human parsing is proposed, in order to improve multi-human parsing in presence of strong occlusions between people in the image.

	RP R-CNN				AIParsing				M2FP			
	mIoU	Δ	AP_{vol}^{P}	Δ	mIoU	Δ	AP_{vol}^{P}	Δ	mIoU	Δ	AP_{vol}^{P}	Δ
Val	60.10	-	59.50	-	60.77	-	60.50	-	68.01	-	62.09	-
D20	58.54	-2.6	56.69	-4.7	58.19	-4.3	57.73	-4.6	66.50	-2.2	59.21	-4.6
D40	57.32	-4.6	54.60	-8.2	55.42	-8.8	56.00	-7.4	64.72	-4.8	58.60	-5.6
D60	55.35	-7.9	49.08	-17.5	50.82	-16.4	51.56	-14.8	63.83	-6.1	57.06	-8.1
D80	46.89	-22.0	48.70	-18.2	46.81	-23.0	50.36	-16.8	57.04	-16.1	57.53	-7.3

Table 3.3: Results for RP R-CNN, AIParsing and M2FP on CIHP instance overlap subsets. The first row (Val) reports values for the whole CIHP validation set. Degree of overlap subsets names have been shortened for visualization purposes (e.g., D20 stands for DoO_20, etc.). Values in Δ columns indicate variations and are expressed as percentages. Percentage symbols are omitted for visualization purposes.

	RP R-CNN				AIParsing				M2FP			
	mIoU	Δ	AP_{vol}^{P}	Δ	mIoU	Δ	AP_{vol}^{P}	Δ	mIoU	Δ	AP_{vol}^{P}	Δ
Val	38.50	-	46.80	-	40.40	-	47.40	-	44.02	-	51.00	-
D20	36.20	-6.0	43.25	-7.6	38.42	-4.9	44.61	-5.9	42.32	-3.9	46.83	-8.2
D40	36.99	-3.9	40.52	-13.4	38.78	-4.0	43.27	-8.7	42.37	-2.9	47.58	-6.7
D60	34.40	-10.6	33.07	-29.3	35.65	-11.8	38.48	-18.8	42.83	-2.7	48.13	-5.6
D80	35.25	-8.4	29.80	-36.3	35.19	-12.9	36.54	-22.9	40.51	-8.0	47.38	-7.1

Table 3.5: Results for RP R-CNN, AIParsing and M2FP on LV-MHP-v2.0 instance overlap subsets. The first row (Val) reports values for the whole CIHP validation set. Degree of overlap subsets names have been shortened for visualization purposes (e.g., D20 stands for DoO_20, etc.). Values in Δ columns indicate variations and are expressed as percentages. Percentage symbols are omitted for visualization purposes.

Chapter 4

CMU Kinoptic-HIS dataset

In order to improve instance-level human parsing performance when dealing with overlapping people in the image, this thesis proposes to exploit multi-view information. In presence of strong occlusions between human bodies, in fact, the biggest challenge for current MHP models, and in particular for top-down approaches, is represented by instance discrimination. Therefore, information provided by multiple views, that frame the same scene from different viewpoints, can be very useful. However, multi-view multi-human datasets with instance-level human segmentation annotations are currently lacking in the literature. Datasets that include multiple people framed from multiple viewpoints are typically designed for 3D human pose estimation and provide only 3D skeletons as ground-truth [45–47]. This brings the necessity of creating an appropriate dataset to enable the development of the new multi-view approach.

This work proposes a technique to produce accurate human instance segmentation annotations from multi-view RGB+D images and 3D human skeletons. In particular, here such method will be applied to the CMU Panoptic Studio dataset. This collects multi-view videos of multiple people interacting and engaging in various activities, providing images rich of overlapping people, which makes it very suitable for the problem dealt with in this thesis.

After an overview on the CMU Panoptic Studio dataset, this chapter describes the devised technique used to recover human instance segmentation annotations from RGB+D images and 3D body skeletons. The ground-truth obtained is very accurate with respect to typical hand-made segmentation annotations.

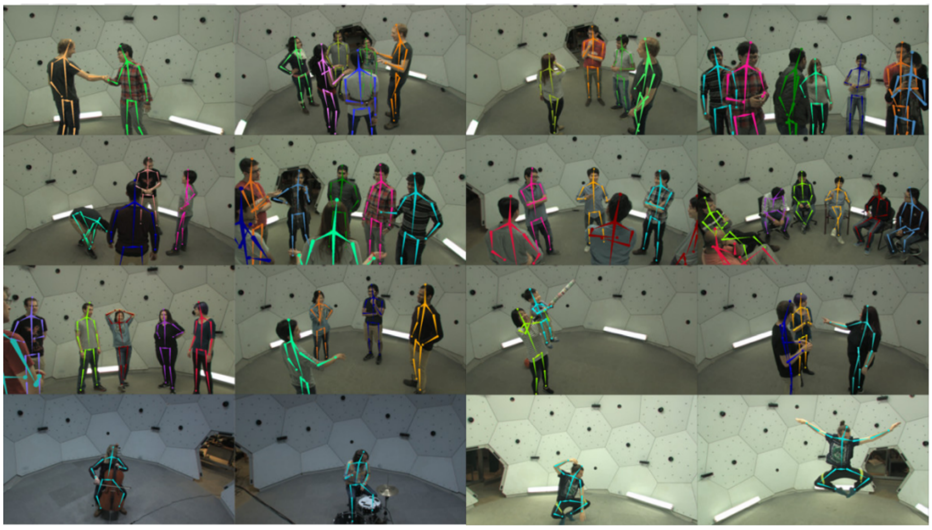


Figure 4.1: Example scenes from CMU Panoptic Studio dataset [15]. In the image also the reprojection of people 3D skeletons.

4.1 CMU Panoptic Studio dataset

CMU Panoptic Studio dataset [15] is a massive multi-view dataset for motion and social interaction capture. It frames different groups of people engaging in various social activities and games (e.g., Ultimatum, Mafia, Haggling, etc.), as well as single people dancing and playing instruments. Some examples of views from the dataset are shown in Figure 4.1. The dataset has been acquired using a multi-camera capture system composed of 480 VGA cameras, 31 HD cameras and 10 Kinect v2 RGB+D sensors, distributed over the surface planes of a geodesic sphere, shown in Figure 4.2. Each camera is calibrated with respect to a common reference frame placed in the center of the dome floor, known as *Panoptic world* reference frame. All calibration data is provided by the dataset.

The dome structure is used to host subjects interactions and to acquire video sequences. The dataset includes more than 198 minutes of videos, amounting to a total of about 154 million frames. The subset of data provided by Kinects is also known as CMU Kinoptic Studio dataset. From this, given the provided calibration parameters, scene point clouds can be generated. This is also referred to as CMU Panoptic Studio PtCloud DB. Figure 4.3 illustrates an example of point cloud generated by merging synchronized depth data for each view.

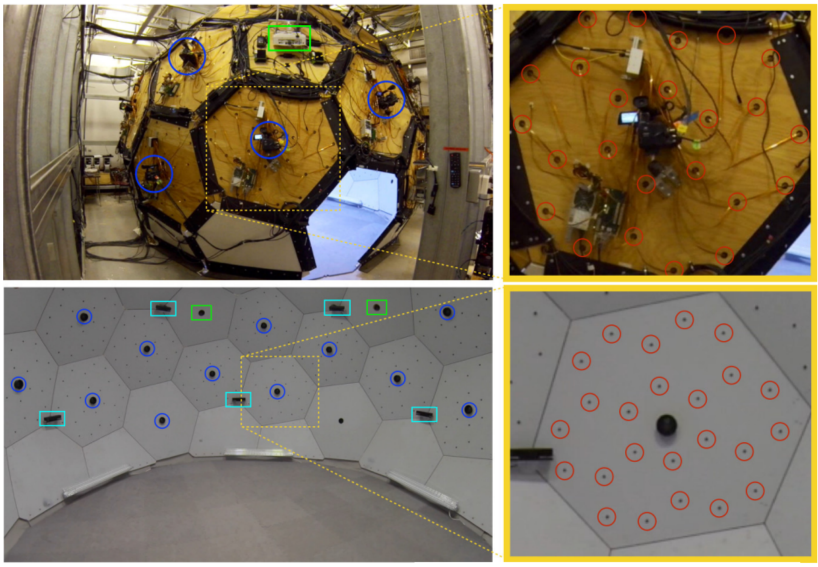


Figure 4.2: CMU Panoptic Studio capture system [15], distributed over the surface of a geodesic sphere having diameter 5.49 meters. VGA cameras are shown as red circles, HD cameras as blue circles and Kinects as cyan rectangles.

4.2 CMU Kinoptic-HIS dataset

This thesis proposes a technique to retrieve accurate human instance segmentation ground-truth from a sequence of multi-view RGB+D videos, annotated with 3D body skeletons. In this case, such method is applied to video sequences from the CMU Kinoptic Studio dataset [15]. This new collection of data is denominated *CMU Kinoptic-HIS*, HIS standing for Human Instance Segmentation. Given the purpose of this dataset, that should include many scenes including overlapping human instances, only videos framing multiple people are considered, in particular: 160224_haggling1, 160226_haggling1, 170407_haggling_a1, 170407_haggling_a2, 170407_haggling_a3 and 160422_ultimatum1. The annotation pipeline will be now described in detail.

4.2.1 Instance-level annotation strategy

The designed annotation technique takes advantage of point clouds and 3D body skeletons to produce human instance segmentation masks from multi-view video sequences. Given video sequences from Kinects sensors, the first step consists in extracting multi-view RGB images. After this, scene point clouds are generated by merging the depth maps captured by the Kinects at the same point in time, that is, in this case, within a time interval of about 15 milliseconds.

In order to illustrate the annotation technique, it is convenient to introduce

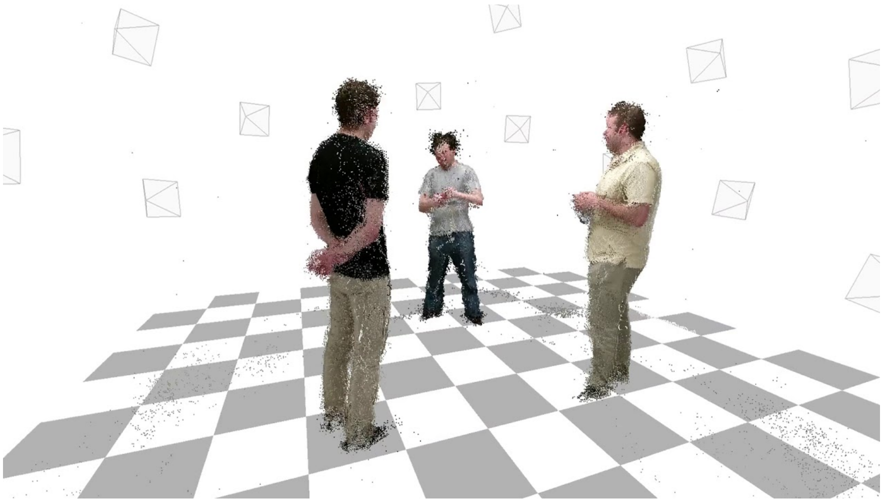


Figure 4.3: Example of point cloud generated from CMU Kinoptic Studio dataset [47].

some details related to CMU Panoptic Studio point clouds and 3D skeletons. Generated point clouds are defined in the so-called *Panoptic world* reference frame, that is, a reference frame placed in the center of the Panoptic dome (Figure 4.2), on the floor. Kinects' RGB cameras reference frames are referred to as *Kinect color* reference frames. To transform the point cloud from the world reference frame into each Kinect color reference frame, two passages need to be performed. First of all, the scale between the two frames needs to be adjusted: 3D coordinates in Panoptic world are defined in centimeters, while in Kinect color frame they are defined in meters. Therefore, there exists a scale factor equal to 0.001 that needs to be applied to 3D points. After this, to get the point cloud in the image reference frame, it is sufficient to apply the rototranslation from Panoptic world reference frame to Kinect color reference frame, defined here as the matrix $T_{panoptic}^{color}$. Ground truth 3D body skeletons provided by the dataset are conveniently defined in Panoptic world reference frame as well. 3D body pose is represented by 19 joints, according to the COCO19 standard, shown in Figure 4.4, defined by the COCO-WholeBody human pose estimation dataset [48].

Point cloud segmentation

In order to obtain instance-level discrimination and instance-indentity consistency between multi-view images, the designed method takes advantage of point clouds and 3D body skeletons. For each point cloud, human bodies are segmented considering the distance between each point and the closest 3D body joint. The closest joint determines the identity of the human instance each point belongs to. In order to make computations more efficient, for each point, the closest body

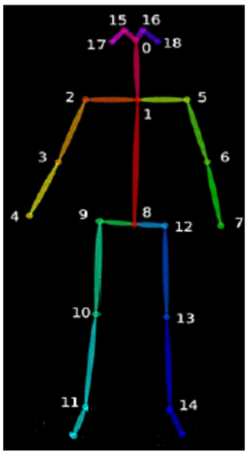


Figure 4.4: COCO19 Keypoint mapping.

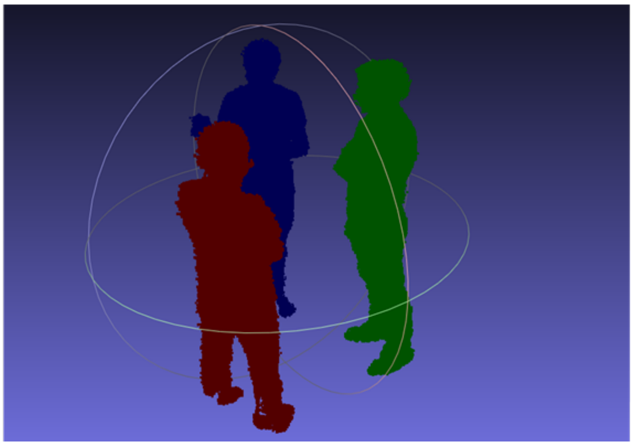


Figure 4.5: Example of segmented and filtered point cloud generated from CMU Kinoptic Studio dataset.

is found by looking in a KDTree [49] of 3D joint coordinates. During this step, the point cloud is also cleaned from noisy and excess points, discarding those whose distance from the closest joint is larger than a threshold value. For a general body joint, the threshold has been empirically set to 50 centimeters. For hands body joints, this threshold is reduced to 15 centimeters, in order to remove 3D points belonging to hand-held objects, that are not of interest. Given that point clouds generated from depth maps fusion are typically noisy, a statistical outlier removal technique is applied, to obtain smoother volumes. In particular, the `statistical_outlier_removal` function from the 3D data processing library Open3D [1], is employed. This function removes points whose distance to their neighbors is too large if compared with the mean distance computed over the whole point cloud. In this case, the neighbor distance for a given point is computed as the average point-to-point distance considering 200 neighbors. Such value was empirically adjusted considering a trade-off between effective outlier removal and excessive point cloud sparseness. The threshold value on point distances depends on the standard deviation of the average distances across the point cloud. Lastly, an input factor determines how aggressive outlier removal will be. Here such value is set to 0.95, to avoid leaving the point cloud too sparse. The result of this first phase is a set of filtered point clouds, that have been segmented and annotated with instance identity. An example is shown in Figure 4.5.

[1] http://www.open3d.org/

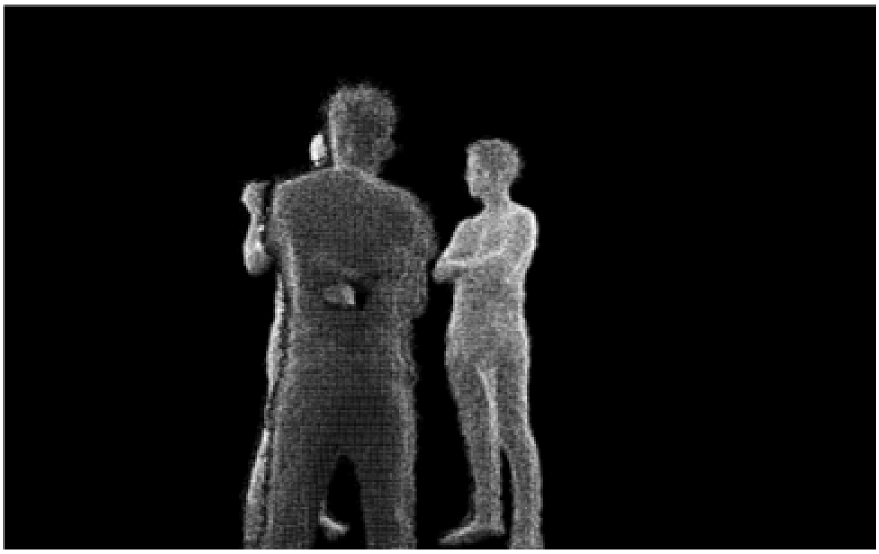


Figure 4.6: Example of result after point cloud projection.

Point cloud projection

As result of our point cloud segmentation step, instance annotations of the people in the scene are obtained for each point cloud. The simplest way to obtain human segmentation masks would be projecting 3D points on each image plane using camera calibration parameters. However, due to the fact that point clouds generated from CMU Kinoptic Studio sequences are quite sparse and noisy, this does not yield satisfactory results. As it can be observed in Figure 4.6, the obtained projections are very sparse, especially for instances close to the camera, there are a lot of holes and contours are not smooth. To solve this issue, this work proposes to retrieve some seed points from sparse projections and use them as segmentation prompts for *Segment Anything Model* (SAM), a transformer-based promptable segmentation system with zero-shot generalization to unknown categories [50]. Given one or more prompts (e.g., points, bounding boxes) hinting at the object in the image that needs to be segmented, SAM is able to produce good quality masks without any additional training or fine-tuning. The proposed annotation approach essentially consists in retrieving good seed points for each projected human instance and using them as prompts for SAM, in order to obtain accurate segmentation masks.

The procedure will be now described in detail. For simplicity, from now on, the annotation pipeline will be described considering a single point cloud and an arbitrary RGB camera. Given the segmented point cloud, each 3D body is transformed from the world reference frame to the camera reference frame, considering the scale factor and $T^{color}_{panoptic}$. After this, points are projected onto the image plane, using intrinsic camera calibration parameters. Seed points for segmentation will be extracted from these body projections. However, since the images will probably frame overlapping human instances, only the actual set of visible, unoccluded points must be considered, in order to avoid using as seeds points that fall onto wrong bodies, leading to inaccurate masks. In particular, if a projected point overlaps with the projection of a body that is closer to the camera (i.e., has smaller depth), it means that the point is occluded and thus, it must be discarded. Considering visible, unoccluded projected points, the goal is to obtain seeds that are evenly scattered over the area that needs to be segmented. To do this, exploiting the K-means algorithm [51], points are clustered and cluster's centers used as seeds. To avoid using too much prompts to segment a very small area of the image, as in the case of occluded humans, the number of clusters computed for each instance depends on the number of points available. If the number of visible points for a body is less or equal to 100, 2 clusters will be computed. If the number of visible points is less or equal to 10000, 4 clusters will be computed. Otherwise, the number of extracted clusters will be 8. These values has been set empirically, after several trials. Using clusters centers as guidance for SAM allows to have seeds evenly scattered over the human instance that needs to be segmented. However there exist some challenging situations in which generated point clouds are extremely sparse in the legs area, producing sparse or truncated body projections, that entirely miss leg points. Experiments show that, especially when segmenting strongly occluded bodies, the absence of seeds for legs leads to incomplete masks. To solve this problem, the set of prompts points is enhanced adding projections for 3D skeletal joints corresponding to knees and ankles (i.e. joints number 10, 11, 13 and 14 in Figure 4.4).

Mask generation

Given the seeds extracted for each instance in the image (i.e., clusters centers and skeletal joints for legs), these are used as prompts for Segment Anything Model, in order to produce masks that will be used as annotations. The steps used by SAM to generate segmentation masks from seed points is shown in Figure 4.7. Despite being a very powerful segmentation model, even when providing appropriate seed points, SAM struggles in accurately segmenting bodies that are largely occluded and tends to group in the output mask portions of overlapping instances. For this reason, this technique proposes to generate masks ordering instances by descending depth, that is, starting from the body that is the farthest away from the camera (i.e., with a larger depth value) and moving on to closer ones. Doing this, occluded bodies will be segmented first and, if generated masks include regions of bodies closer to the camera, these will be overwritten. After the segmentation mask for a given instance is produced, it is fed again to the SAM network for refinement. During this step, seed points for all the other human instances in the image are provided as well, labelled as belonging to background. This has the effect of helping SAM identifying the actual target human that needs to be segmented and leads to more accurate masks. Figure 4.8 depicts some sample annotation masks produced using the proposed strategy on the CMU Kinoptic Studio dataset. As it can be observed, human instance segmentation masks show good quality even when bodies are largely occluded.

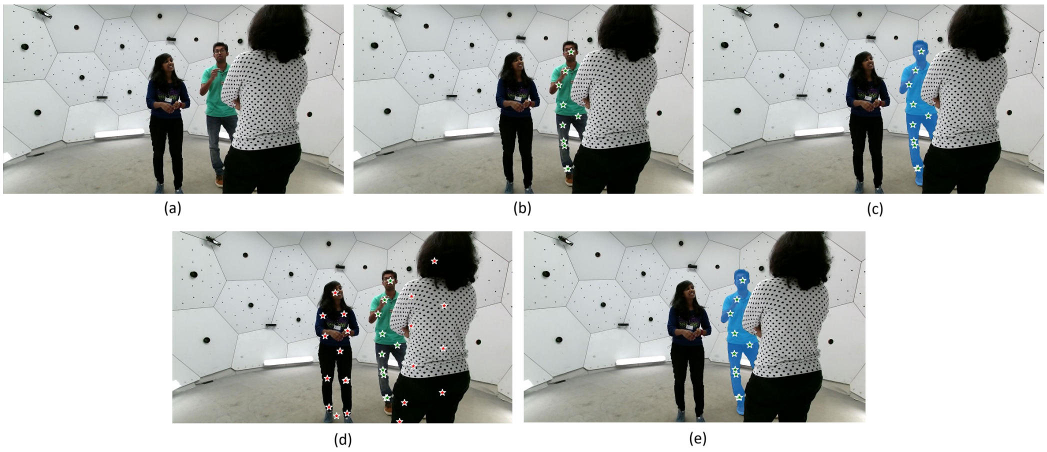


Figure 4.7: Segment Anything Model segmentation steps. RGB image (a). Seed points (b). Mask generated from seed points (c). Seed points used for refinement, red points indicate regions that should be considered as background (d). Final refined mask (e).

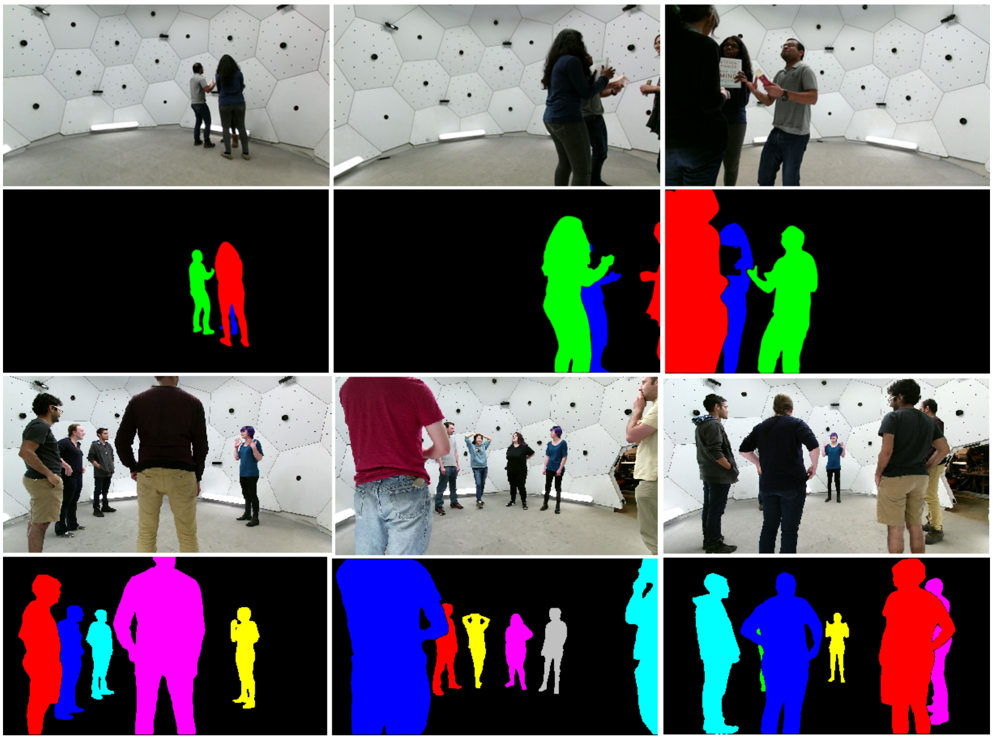


Figure 4.8: Example of annotation masks generated for CMU Kinoptic Studio dataset.

Chapter 5

Multi-view instance-guided multi-human parsing

This thesis focuses on enhancing multi-human parsing when dealing with strong occlusions between people. Information from multiple views is very strategic in these scenarios: people that appear as overlapped in an image, will appear as separated if framed from different viewpoints. Following this intuition, this work proposes to exploit multi-view information to improve segmentation results in case of occlusions between people. Taking inspiration from Caliskan et al. [52], that leverage multi-view human shape consistency to guide 3D human body reconstruction, multi-view human instance segmentation ground-truth is exploited to guide the multi-human parsing task, improving how it recovers human bodies when heavily occluded. This is done according to a novel learning framework which aims at enhancing the performance of state-of-the-art multi-human parsing techniques, injecting 2D and 3D instance-level information about human bodies. Consider for example a multi-human parsing architecture that has already been trained on a specific multi-human parsing dataset (e.g., CIHP [13]). The proposed method consists in performing a fine-tuning procedure using a multi-view dataset with human instance segmentation ground-truth, as the one produced in Chapter 4. This provides a weak supervision for multi-human parsing, as the exploited annotations represent the original task at much coarser granularity. Nevertheless, typical multi-human parsing annotations include both instance discrimination and body parts labels, while human instance segmentation provides distinction between background and foreground only. The fine-tuning is guided by two classes of loss functions:

- *single-view loss functions on human instance segmentation*, which leverage

2D human instance information to encourage the network to segment whole bodies and to effectively identify the target instance when severely occluded;

- *multi-view loss function for instance identity and body part prediction consistency*, that exploits knowledge about 3D human instances to promote instance identity disambiguation and to enforce consistent body parts predictions between different views of the same scene.

The learning framework exploiting single-view loss functions only will be referred to as Instance-Guided Multiple Human Parsing (IG-MHP), while the approach including both single-view and multi-view loss contributions will be referred to as Multi-View Instance-Guided Multi-Human Parsing (MVIG-MHP). In the following sections, details about each method will reported and described.

5.1 Instance-guided multi-human parsing

Considering a multi-human parsing architecture trained on a multi-human parsing dataset (e.g., CIHP), the obtained model is fine-tuned exploiting an auxiliary human instance segmentation dataset, such as the CMU Kinoptic-HIS dataset described in Section 5.2. Human instance supervision is used to guide multi-human parsing with the aim of enhancing human instance recovery on challenging images in which bodies are heavily overlapped.

After the fine-tuning, the focus is again on the main task and the network is evaluated on the target multi-human parsing dataset. For this reason, special attention has to be paid to the fine-tuning procedure: learning on a new dataset using weaker body part-agnostic supervision could disrupt important knowledge about accurate parts prediction. For example, the CIHP dataset includes a wide range of images framing people appearing in different poses, viewpoints, scales and resolutions, whereas images from the CMU Kinoptic-HIS dataset have all been taken in the same place and have very similar backgrounds. People pose does not change widely throughout the dataset and the same groups of people appear in many images. Moreover, some CIHP semantic categories such as glove, dress or skirt are not represented. Trying to avoid the detrimental effects that this domain gap could have on multi-human parsing accuracy, the proposed fine-tuning procedure does not affect the whole network. Taking into consideration top-down architectures (e.g., AIParsing) weights related to feature backbone and detection head should be kept frozen and not modified. Leaving the backbone unconstrained, in fact, would most certainly lead to features alterations, which would then translate in poorer performances when going back to CIHP for evaluation. Similarly, given the different degree of variety in people scale and pose between the datasets, the detection head would suffer strong modifications trying to adapt to the new data, which could be disruptive when dealing with the original dataset again.

5.1.1 Single-view loss functions for human instance segmentation

To take advantage of the human instance segmentation ground-truth as weak supervision for multi-human parsing, the fine-tuning needs to be guided by proper losses. The proposed loss functions take inspiration from the ones proposed for AIParsing architecture in [30].

Parsing predictions for largely occluded instances tend to be partial and include a lot of holes. Human instance segmentation is, thus, exploited as auxiliary task to improve the ability of the multi-human parsing network to recover complete and accurate human parsing maps for largely occluded people. To do this, fine-tuning is guided by two main loss functions, $L_{instance\_fg}$ and $L_{instance\_seg\_miou}$. Consider for example a top-down multi-human parsing architecture. Given a region of interest extracted by the detection head, the loss function $L_{instance\_fg}$ guides the parsing sub-network to accurately segment the correct human instance: body parts belonging to overlapping bodies that fall into that same region of interest should be disregarded and assigned to background. $L_{instance\_fg}$ represents the standard cross-entropy loss between human segmentation predictions and ground-truth. The considered network, however, has been trained to predict body parts for each human instance, therefore, obtaining just human instance segmentation predictions is not straightforward. This thesis proposes to consider the union of body parts parsing maps, in order to obtain full body predictions. Specifically, body parts predictions on a single region of interest are fused into a single map by choosing the maximum output value for each pixel. In other words, considering human and background categories, the probability that each observation x is classified as human is given by:

$$p_h(x) = \max\{p_i(x) : i \in C\} \tag{5.1}$$

where C represents the set of CIHP body parts categories, for example. To further improve the quality of the human instance segmentation obtained by following $L_{instance\_fg}$, $L_{instance\_seg\_miou}$ directly optimizes the mean Intersection-over-Union between predicted and ground-truth human segmentation masks, exploiting Lovász-Softmax formulation [43]. Since mean Intersection-over-Union essentially quantifies the overlap between predictions and ground-truth, $L_{instance\_seg\_miou}$ has the effect of punishing the network when producing incomplete human bodies segmentation masks (e.g., internal holes in the segmentation mask). Combining loss contributions given by $L_{instance\_fg}$ and $L_{instance\_seg\_miou}$, encourages prediction of accurate parsing maps for the correct human instance, recovering accurate body outlines and filling holes. An additional refinement loss function, suitable to top-down approaches, $L_{instance\_box\_miou}$ can be also be introduced to consider the mean Intersection-over-Union for the predicted human instance map within the corresponding detected bounding box. This is useful to filter out poor quality candidate human regions, similarly to what is done in [30]. All considered, the proposed fine-tuning is driven by a loss function composed by three terms:

$$L_{IG-MHP} = L_{instance\_fg} + L_{instance\_seg\_miou} + L_{instance\_box\_miou} \tag{5.2}$$

The network will predict the multi-human parsing task on the human segmentation auxiliary dataset, with a particular encouragement towards producing complete human instance parsing maps with accurate boundaries.

5.2 Multi-view instance-guided multi-human parsing

Building upon instance-guided multi-human parsing, based on single-view loss functions only, multi-view instance-guided multi-human parsing adds a multi-view loss term that exploits knowledge about 3D human instances to encourage instance identity discrimination and body parts prediction consistency. A main challenge with occlusions between people in a RGB image is due to an intrinsic ambiguity between their bodies, as they are projected on the same 2D plane, one overlapped with the other. In this case, it becomes difficult to distinguish between the two using an RGB image only, as their body parts are intersecting. Exploiting 3D information allows to resolve this ambiguity, as human bodies in the 3D space are more easily separated. The idea of resorting to multi-view information to enhance multi-human parsing on strong occlusions between human instances comes from this intuition. When the same scene is framed from multiple points of view, ambiguity between human instances is easily removed and associating body parts to the correct person becomes very simple. At the same time, between multiple views there must be coherence, as the framed 3D scene is identical. In other words, if a certain 3D point belongs to a given human instance and body part, this has to be true from each viewpoint, that is, for each view. This can be referred to as multi-view consistency.

5.2.1 Multi-view loss function for instance identity and body part prediction consistency

To improve disambiguation between overlapped human instances and to enforce consistency in body parts predictions between multiple views, the proposed learning framework introduces a loss contribution on multi-view images. This is done exploiting the 3D point clouds provided with human instances annotations made available by the annotation procedure described in Section 4.2.

Consider multiple adjacent views I_i with i from 1 to $\mathcal{N}$, $\mathcal{N}$ at least equal to 2, and a 3D human body point P. The corresponding forward-projection on view I_i is referred to as FP_i. Multi-view consistency loss encourages coherent instance identity predictions between point projections of the same 3D point P in multiple views. This means that if point P is associated to a given human instance by 3D ground truth, each projection in $\{FP_i\}$ should be associated to the same human instance as well. Multi-view consistency on instance identity, aims at improving separation and body part matching accuracy between overlapped human bodies.

Similarly, 3D human instance information can be exploited to enforce consistency in body part prediction between multiple views. Note that the losses designed for single-view approach presented in Section 5.1 relies just on 2D human instance information to guide the network, leaving the body part prediction without supervision. When using multi-view information to guide the network fine-tuning, predictions of body parts from each view should be consistent with each other; by imposing such consistency at the loss level, weak supervision of body parts can be introduced. To enforce body part prediction consistency across different views, this thesis takes inspiration from Antonello et al. [53], that fuse pixel-level predictions from multiple views to refine single-view semantic segmentation. Considering all forward-projections $\{FP_i\}$ of the same 3D human body point P on multiple views, the optimal body part label for P can be estimated aggregating all contributions from the $\mathcal{N}$ views. In particular each forward-projection participates with the body part label c predicted by the network to be fine-tuned and with the label confidence $p(FP_i|c)$ given by network. The label predicted by the most views, with the highest aggregated confidence, is selected as optimal for the 3D point P and propagated on all views. In other words, the optimal label c^* for a point P is given by:

$$c^* = \underset{c \in \mathcal{C}}{\operatorname{argmax}} \sum_{i=1}^{\mathcal{N}} p(FP_i|c) \tag{5.3}$$

where $\mathcal{C}$ is the set of body part labels considered by CIHP, for example. This is done considering the cross-entropy loss function between the optimal label obtained by aggregating contributions, and the predicted label, for each view.

5.2.2 Multi-view forward projection

The forward projection to map information from the 3D instance ground truth to the multi-views images is a crucial step in the computation of the multi-view loss function during the network fine-tuning. As described in Section 5.2.1, in such operation each 3D point of each human instance is projected on different views of a same scene, obtaining a set of points on which instance and body part consistency can be imposed. However, few different considerations should be kept in mind in implementing such operation. Firstly, in case of many people in the scene, they can appear as occluded from a certain viewpoint. In such cases, 3D points belonging to occluded people should not be projected. Moreover, projecting all the 3D points of all human instances can take very long time and computational resources with a negative impact on the fine-tuning feasibility. In order to deal with such problems, in this thesis the forward projection is implemented introducing two thresholds: one related to the distance of each human instance from the camera, and one related to the number of 3D points considered in the projection. When projecting 3D points onto a 2D plane, there exist ambiguity, as more than one 3D point is projected onto the same 2D point. This represents a huge problem when considering contributions for label fusion, as the forward projection of a 3D point belonging to an occluded region could contribute with the wrong label. Considering, for example, a human instance framed from the side, 3D points belonging to the occluded arm should not be included in label fusion, as they would contribute with a wrong part category. To avoid such issues, a 3D body point is considered and projected only if their distance from the body point which is closer to the camera is lower than a certain threshold. Such threshold on points distance, as well as the maximum number of points considered for the projections are set empirically.

5.3 Training details and best model selection criteria

To describe the proposed framework, this thesis will consider state-of-the-art top-down architecture AIParsing [30] as starting point, CIHP [13] as target multi-human parsing dataset and CMU Kinoptic-HIS as human instance segmentation ground-truth used to carry out the fine-tuning. The training set used to fine-tune the model is composed by 1360 images, taken from four different sequences, namely *160224_haggling1*, *160226_haggling1*, *170407 _haggling_a1* and *160422_ultimatum1.* While *haggling* sequences frame up to 3 people per image and *ultimatum* sequences frame up to 8 people per image, the 40% of the training dataset is drawn from *160422_ultimatum1*, while the remaining 60% is drawn from the other ones. As already explained, CMU Kinoptic IS images do not present large variance in background, people appearance, pose and scale, leading to the decision of using a modest number of images for training.

To monitor the training process and avoid overfitting on the training data, a set of 1000 images is used as validation set. These are taken from two different sequences, without overlap with those that have been selected for the training set. In particular, images from *170407_haggling_a2* and *170407_haggling_a3* are employed. While performances on the validation set are usually taken into account in order to select the best model during training, this would not be significant in this case, as performances are evaluated on the target dataset CIHP. To choose the best model, the network is trained for a fixed number of epochs, after which each model is evaluated on a subset of the CIHP target dataset, specifically, on CIHP_DoO_60, that is, the subset of CIHP images that frame people having degree of overlap at least equal to 0.6. Choosing the best model on such set of images is consistent with the main focus of this work, which is contrasting challenges given by strong occlusions. The mean Intersection over Union, representing body parts segmentation quality, in particular, is selected as measure of goodness for the obtained models. The set of network parameters that achieve the higher mIoU on the validation set, are deemed as the best ones. The mean Intersection over Union is selected to choose the final model as it gives a good representation of the model behaviour on the target dataset. Being body-part-aware, an improvement on this metric indicates that the overall body parts segmentation quality has been enhanced, and thus, that the fine-tuning on the human instance segmentation domain has been effective and, most importantly, did not disrupt most of the knowledge learned by the model on the original MHP dataset.

To examine the results of the fine-tuning carried out on the network, along with the typical multi-human parsing metrics illustrated in Section 3.1, (e.g., mean Intersection over Union, Average Precision based on Part, Probability of Correct Parts), evaluation metrics measuring human segmentation quality must also be considered. In particular, given the loss functions employed to drive the network, the mean Intersection over Union on human instance segmentation, here referred to as mIoUh, provides a good representation of the effectiveness of this stage. In particular, if such part-agnostic mIoU improves and part-aware mIoU improves as well, it means that the fine-tuning was successful: foreground human segmentation was enhanced and the network preserved its ability to predict accurate body parts, despite seeing coarser ground-truth.

Chapter 6

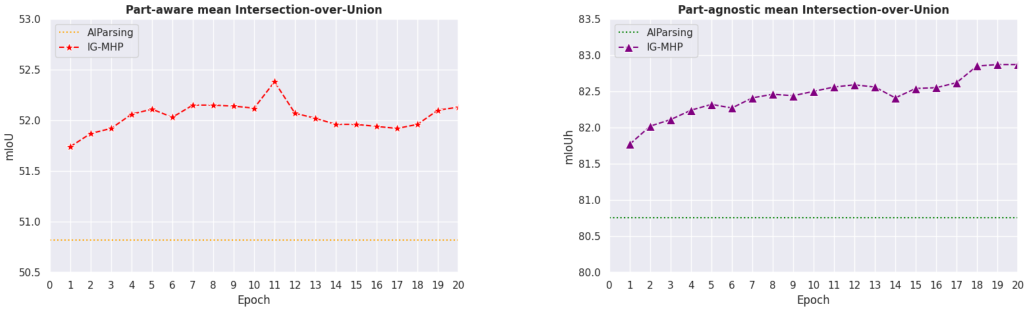
Experimental results for instance-guided multi-human parsing

In this chapter, the instance-guided multi-human parsing approach presented in Section 5.1 will be evaluated through different experiments, aiming at improving the performance of a state-of-the-art architecture, such as AIParsing [30], in the challenging scenario of heavily overlapped people. Instance-guided multi-human parsing, here referred to as IG-MHP, exploits single-view human instance segmentation loss functions in order to guide the multi-human parsing task. The experimental validation of the proposed approach presented in this chapter will consider state-of-the-art top-down architecture AIParsing as starting point, CIHP [13] as target multi-human parsing dataset and CMU Kinoptic-HIS, presented in Section 4.2, as human instance segmentation ground-truth used to carry out the fine-tuning. In accordance with the learning framework introduced, model performances are measured on the multi-human parsing dataset used to pre-train the considered network (i.e., CIHP). Different subsets of images with increasing degree of overlap are considered (i.e., CIHP_DoO_20, CIHP_DoO_40, CIHP_DoO_60, CIHP_DoO_80), in order to highlight the advantage achieved on the specific scenario of interest. An analysis taking into account different hyperparameters, such as learning rate, will be then carried out, in order to optimize the proposed method. To measure performances on the target task, both global and instance-level evaluation metrics will be considered, as defined in Section 3.1. Mean Intersection-over-Union (mIoU), pixel accuracy (pix_acc) and mean pixel accuracy (mean_acc) on CIHP semantic categories will be examined to evaluate body parts segmentation quality, while Average Precision based on Part (AP^P_{50}, AP^P_{vol}) and Probability of Correct Parts (PCP) will provide an indication on the instance-level part segmentation results. Mean Intersection-over-Union on human instance segmentation, here denoted as mIoUh, will also be considered. When mentioning mean Intersection-over-Union measures, to avoid ambiguity, the mIoU on body parts will sometimes be referred to as part-aware mIoU, while mIoU on human instance segmentation will be sometimes be referred to as part-agnostic mIoU. For simplicity, in the following experiments, the model prior to fine-tuning will be referred to as AIParsing, since it represents the multi-human parsing method as proposed by Zhang et al. in [30]. This will also be the baseline for the presented experiments.

6.1 Instance-guided multi-human parsing on overlapping human instances

The performances of the proposed instance-guided multi-human parsing method on overlapping human instances will be now compared with results obtained from state-of-the-art network AIParsing, used as baseline. As already mentioned in Chapter 2, AIParsing was originally trained for 75 epochs on the CIHP training set with image batches of size 8, using an initial learning rate equal to 0.005, decreased by a factor of 10 each 50 and 65 epochs, respectively. Here AIParsing is fine-tuned on CMU Kinoptic-HIS exploiting human instance segmentation loss functions only. The batch size is set to 8 images, to be consistent with the original AIParsing implementation. The learning rate for this experiment is set to 5e-5, that is the learning rate used by the network for the last epochs of training. Due to time constraints for carrying out the experiments, the network is fine-tuned for a fixed number of epochs, 20 in this case, and the best model is selected according to the part-aware mIoU measure on a subset of the target dataset CIHP, namely CIHP_DoO_60, as already described in Section 5.3.

Figure 6.1a shows the trend of the part-aware mean Intersection-over-Union computed on CIHP_DoO_60 over the 20 fine-tuning epochs, with respect to the value that is reached for the same metric by the baseline method AIParsing. It can be observed that the proposed approach leads to an improvement after one epoch of training, proving the effectiveness of the proposed method. Such result highlights an enhancement in body parts segmentation quality in presence of human instances that are largely occluded. Recalling the meaning of the part-aware



(a) Part-aware mean Intersection-over-Union for instance-guided multi-human parsing method.

(b) Part-agnostic mean Intersection-over-Union for instance-guided multi-human parsing method.

Figure 6.1: Part-aware mean Intersection-over-Union and part-agnostic mean Intersection-over-union for instance-guided human parsing method.

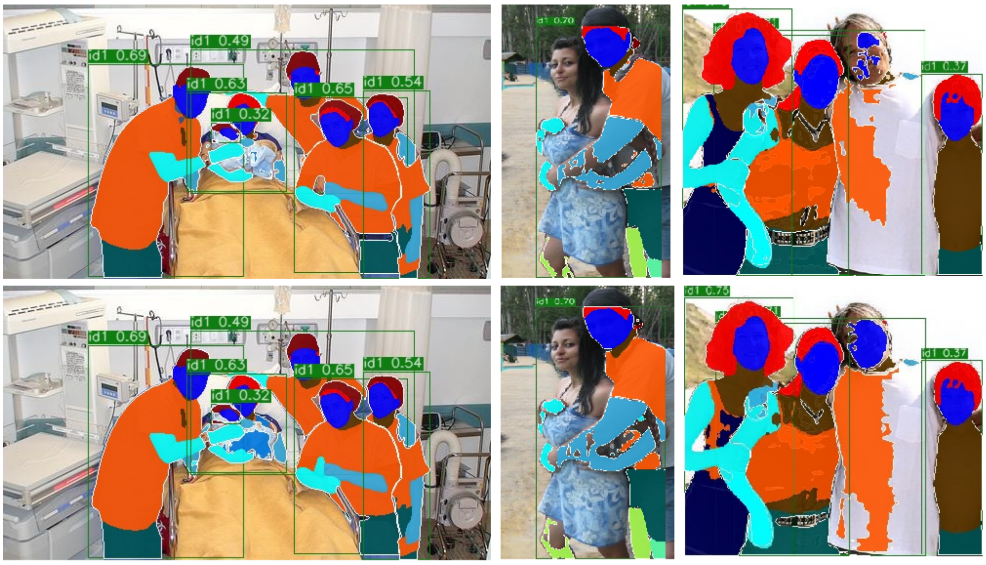


Figure 6.2: Comparison between multi-human parsing given by AIParsing (top row) and single-view method (bottom row).

mIoU evaluation metric, which takes into account the Intersection-over-Union for each semantic class, an improvement with respect to the baseline model means that the network is effectively learning the auxiliary human instance segmentation task, leading to a more accurate segmentation of whole bodies and, at the same time, the valuable knowledge on body part discrimination, acquired on the multi-human parsing dataset, is preserved. A qualitative comparison between AIParsing and IG-MHP is illustrated in Figure 6.2, highlighting the contribution of the proposed approach.

Confirming the effect of IG-MHP on human instance segmentation, Figure 6.1b shows the evolution of the part-agnostic mean Intersection-over-Union over the fine-tuning epochs. It is clear that the loss terms and the new data are successfully driving the fine-tuning. Moreover, the learned cues do not disrupt the ability of the network on the target task but actually, performance is enhanced. It is worth noticing that values obtained for the part-agnostic mean Intersection-over-Union are much higher than those obtained for the part-aware mean Intersection-over-Union, as shown in Figures 6.1a and 6.1b. This is due to the fact that, while the first one only considers two classes (i.e., human and background), the second considers 20 different categories, and therefore, it is much more difficult, but also much more significant, to obtain improvements. This validates the choice of selecting the best model according to the part-aware mIoU, instead of looking at mIoUh. Examining the mIoU trend over the epochs, the metric reaches its peak after approximately 10 epochs of training and the best model is obtained after 11 epochs.

For a more in-depth analysis of the performance of the instance-guided method on overlapping human instances, the best model found was evaluated also on the other subsets of images with overlapped people, namely CIHP_DoO_20, CIHP_DoO_40 and CIHP_DoO_80. Results are shown in Table 6.1. It can be noticed that, the more the degree of overlap of the considered images increases, the more the proposed method brings a relevant advantage on body parts segmentation quality. It can be, however, observed that, despite bringing improvements on CIHP subsets including huge occlusions (i.e., CIHP_DoO_60 and CIHP_DoO_80), when considering also images in which occlusions are not as severe, the advantage gained is lost. This could be related to the fact that during the fine-tuning process on the CMU Kinoptic-HIS dataset, body parts prediction is left almost unsupervised. Indeed, body parts predicted by the network are not corrected by a dedicated loss function during the fine-tuning; the network learns to refine them in order to improve the instance segmentation, leading to slight changes to predicted classes as training proceeds. When facing very challenging scenarios, such as largely occluded people, the proposed approach is able to recover more accurate human instances, counter-balancing this deterioration in body part prediction. When evaluated on a larger set of images on which the baseline method already performs satisfactorily (e.g., CIHP_DoO_20), this effect becomes visible. Nevertheless, this proves that the scenarios on which the baseline struggles are represented by sets of images having a degree of overlap of at least 40%. Therefore, given the focus of this work, these are the sets of data on which testing the proposed method becomes very relevant. For this reason, when analysing behaviour in case of overlapping people, CIHP_DoO_40, CIHP_DoO_60 and CIHP_DoO_80 will mainly be

	Model	mIoU	pix_acc	mean_acc	mIoUh	AP_{50}^{P}	AP_{vol}^{P}	PCP
DoO_20	AIParsing	**58.19**	**88.68**	68.32	88.41	**71.64**	**57.73**	**68.55**
	IG-MHP	57.70	88.55	**68.49**	**88.85**	60.84	55.98	64.81
DoO_40	AIParsing	55.42	87.57	64.95	86.52	**68.75**	**56.01**	**63.24**
	IG-MHP	**55.46**	**87.69**	**65.78**	**87.49**	65.81	54.31	62.16
DoO_60	AIParsing	50.82	84.84	59.47	80.75	**61.09**	**51.56**	**57.72**
	IG-MHP	**52.38**	**85.41**	**62.02**	**82.59**	57.55	49.83	56.20
DoO_80	AIParsing	46.81	83.08	58.25	79.21	**58.43**	**50.36**	**55.59**
	IG-MHP	**47.42**	**83.76**	**59.51**	**81.13**	55.92	48.63	54.09

Table 6.1: Overall performance of IG-MHP on CIHP_DoO_20, CIHP_DoO_40, CIHP_DoO_60 and CIHP_DoO_80 subsets, respectively indicated as D20, D40, D60 and D80 for visualization purposes.

considered.

Looking at results in Table 6.1, instance-level metrics such as AP_{50}^{P} and PCP decrease. This is indeed another symptom of the fact that body part prediction is being affected, and such metrics attribute significant weigh to part accuracy.

6.1.1 Analysis on learning rate choice

Learning rate plays a critical role in network fine-tuning. As already mentioned, features learned by the model during the pre-training phase on CIHP are extremely important. If the learning rate is too high, pre-trained network's weights will change dramatically: the model will forget how to properly segment body parts and start to predict a single class for whole human instances. By using a rather small learning rate, the network can slowly adapt to the new data, without moving too much from the optimal solution found on the starting dataset. However, if the learning rate is too small, the network is not encouraged to learn new insights and the fine-tuning on the new data is not exploited to its full potential. This considered, for this experiment the learning is increased to a value equal to 3e-4, in order to evaluate if the previous learning rate was already a good choice or if increasing it leads to better performances without degrading task-specific knowledge.

The evolution of the part-aware mean Intersection-over-Union during fine-tuning is shown in Figure 6.4a. The behaviour of this metric, in this case, is quite different to the one observed for the first experiment. There are more oscillations and less plateaus in the observed values, which is very consistent to the fact that the learning rate employed is higher and thus, the model loss function makes bigger leaps during optimization. A comparison between mIoU

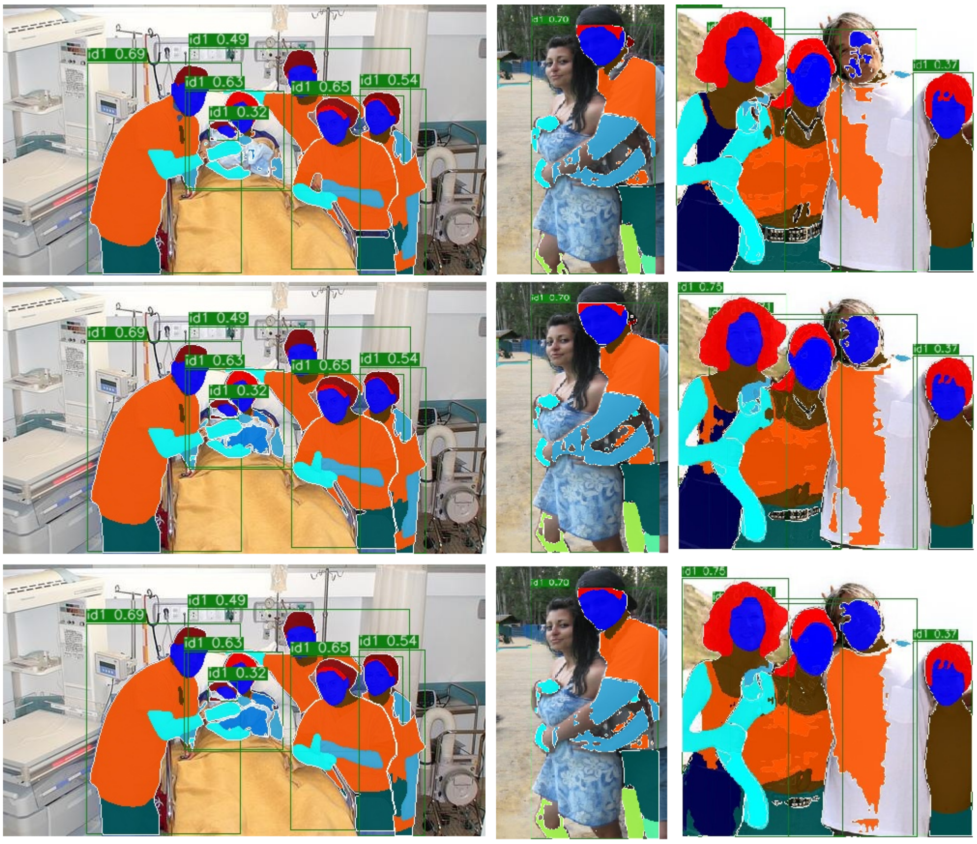


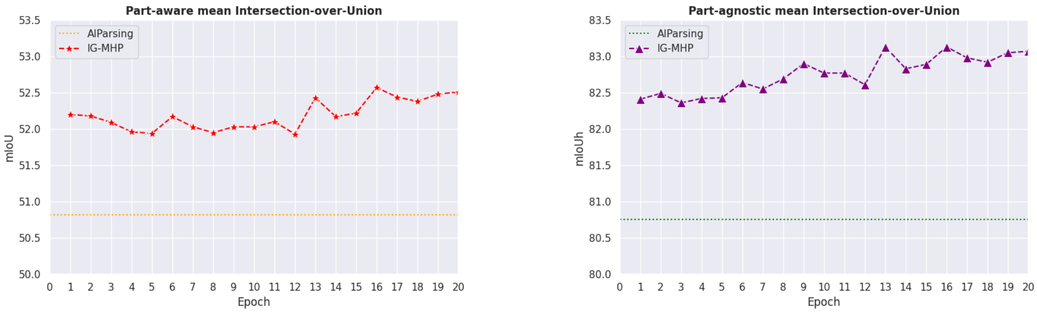
Figure 6.3: Comparison between multi-human parsing results given by the baseline AIParsing (top row) and IG-MHP with learning rate equal to 5e-5 (middle row) and IG-MHP with learning rate equal to 3e-4 (bottom row).

evolution for learning rate equal to 3e-4 and for learning rate equal to 5e-5 (i.e., first experiment) is illustrated in Figure 6.5.

The best model is obtained after 16 epochs and reaches a higher accuracy in terms of part-aware mIoU with respect to the previous experiment, showing that the approach benefits from a higher learning rate. The network is more encouraged to adapt to the new data, as testified by the growth of the part-agnostic mean Intersection-over-Union during fine-tuning (Figure 6.4b) and the performance improvement on overlapped human instances is more relevant, as it can be observed from Table 6.2. At the same time, however, the effect of the body part precision degradation becomes visible even on CIHP_DoO_40. This confirms that there exists an important trade-off between the advantage reached on severely occluded bodies, given by the new knowledge learned during the fine-tuning, and body part precision. The more the network adapts to the new data, depending on fine-tuning duration and learning rate, the more the task learned during the pre-training on multi-human parsing undergoes some changes. This is a straight consequence of the higher coarseness of the annotations used for fine-

	Model	mIoU	pix_acc	mean_acc	mIoUh	AP^P_{50}	AP^P_{vol}	PCP
DoO_40	AIParsing	**55.42**	87.57	64.95	86.52	**68.75**	**56.01**	**63.24**
	IG-MHP	55.23	**87.62**	**65.63**	**87.68**	63.52	53.22	61.28
DoO_60	AIParsing	50.82	84.84	59.47	80.75	**61.09**	**51.56**	**57.72**
	IG-MHP	**52.57**	**85.51**	**62.49**	**83.13**	56.33	48.75	55.26
DoO_80	AIParsing	46.81	83.08	58.25	79.21	**58.43**	**50.36**	**55.59**
	IG-MHP	**47.61**	**83.95**	**60.03**	**81.97**	53.09	47.54	53.04

Table 6.2: Overall performance of instance-guided multi-human parsing on CIHP_DoO_40, CIHP_DoO_60 and CIHP_DoO_80 using learning rate equal to 3e-4.



(a) Part-aware mean Intersection-over-Union for instance-guided multi-human parsing method.

(b) Part-agnostic mean Intersection-over-Union for instance-guided multi-human parsing method.

Figure 6.4: Part-aware mean Intersection-over-Union and part-agnostic mean Intersection-over-union for instance-guided multi-human parsing method with learning rate equal to 3e-4.

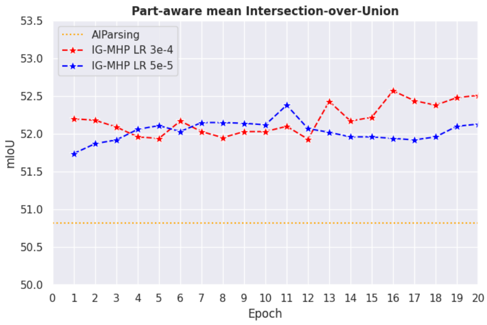


Figure 6.5: Comparison between part-aware mean Intersection-over-Union evolution using learning rate equal to 3e-4 and 5e-5 during instance-guided fine-tuning.

tuning and of the domain gap between data used for training and for evaluation.

A qualitative comparison on results illustrated in Figure 6.3, highlights the contribution of the instance-guided method with a higher learning rate. Performance on occluded bodies is visibly enhanced with respect to the baseline as the model is able to recover significant portions of bodies that AIParsing misses.

6.1.2 Analysis on data augmentation strategies

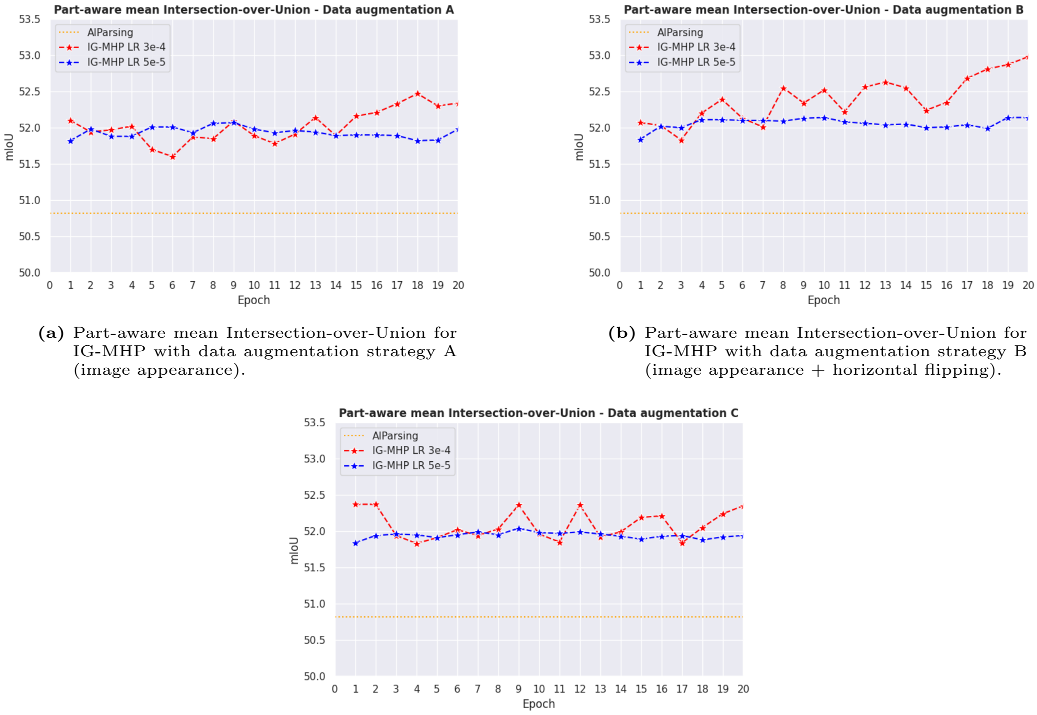
As highlighted by previous analyses, there exits a quite huge domain gap between CIHP and CMU Kinoptic-HIS datasets. Samples from the new set of data present less variance in appearance, pose and most of all, scale. Reducing this gap could help improving performances on the target dataset. To test this hypothesis, three experiments are carried out, adding different data augmentation strategies to fine-tuning. In particular, data augmentation on image appearance, flipping, scaling and cropping is applied. Scale and resolution with which human bodies appear in the image is one of the biggest discrepancies between the two datasets. For this reason, all the experiments include image resizing as data augmentation strategy. This is performed on the input images with a probability equal to 0.3.

A) Image appearance data augmentation

All images from CMU Kinoptic-HIS are taken in the same place, that is, the Panoptic Studio dome, as illustrated in Section 4.2. For this reason, all images present very similar backgrounds. Considering this, random background replacement is adopted to introduce higher variety and applied on almost every input image (i.e., with probability 0.98). Background images have been collected from Google and frame both outdoor and indoor scenes, to resemble CIHP images as much as possible. Additionally, gaussian blur and color jitter are also applied, with probability equal to 0.3. Blur varies with kernel size from 5x5 to 11x11 and sigma from 0.1 to 2. For color jitter, only variations in brightness and contrast are introduced while hue is left untouched.

B) Adding image flipping data augmentation

Overcoming the fact that people pose does not vary much throughout the images used for fine-tuning is not an easy task. Image flipping could be of some help, introducing some variety in body configuration. For this, to the data augmentation on image appearance used in experiment A, random image horizontal flipping is



(a) Part-aware mean Intersection-over-Union for IG-MHP with data augmentation strategy A (image appearance).

(b) Part-aware mean Intersection-over-Union for IG-MHP with data augmentation strategy B (image appearance + horizontal flipping).

(c) Part-aware mean Intersection-over-Union for IG-MHP with data augmentation strategy C (image appearance + horizontal flipping + image cropping).

Figure 6.6: Part-aware mean Intersection-over-Union for IG-MHP with different data augmentation strategies, for the tested learning rates.

applied. In practice, with probability 0.3, the network will see sample images in which human bodies appear upside down.

C) Adding body scale, truncation and occlusion data augmentation

For this last experiment, random image crop is also added. It is performed with probability 0.3 and introduces occlusions, body parts truncation and stronger variations in terms of scale, with respect to random image resize.

For each combination, best model selection is done as for previous experiments. Figure 6.6 shows the trend of part-aware mean Intersection-over-Union for each of the three strategies, comparing models with the different learning rates tested.

Table 6.3 shows global-level and instance-level evaluation metrics on CIHP_DoO_60 for the best models obtained incorporating data augmentation techniques. Learning rate equal to 3e-4 is the most effective, for all experiments. The data augmentation strategy which brings the largest improvement on overlapping images is the one illustrated in B). This combines variations related to

	mIoU	pix_acc	mean_acc	mIoUh	AP^P_{50}	AP^P_{vol}	PCP
AIParsing	50.82	84.84	59.47	80.75	**61.09**	**51.56**	**57.72**
IG-MHP	52.57	85.51	**62.49**	83.13	56.33	48.75	55.26
IG-MHP-A	52.47	85.51	61.81	83.00	56.99	49.19	55.94
IG-MHP-B	**52.98**	**85.59**	62.34	**83.18**	57.39	49.42	55.94
IG-MHP-C	52.22	85.36	61.68	82.35	58.91	50.33	57.03

Table 6.3: Performances on CIHP_DoO_60 subset for the best models found for each data strategy.

appearance and body configuration and leads to a very significant advantage in terms of part-aware mean Intersection-over-Union, with respect to the AIParsing baseline. It is worth noticing that using strategies A and C, despite improving with respect to the pretrained network, the performance is slightly worse if compared to the model which does not employ any data augmentation. This is an effect of the reduced gap between the two datasets. This leads to a smaller improvement on overlapped instances but body part precision is less affected. This is confirmed by instance-level evaluation metrics, which improve with respect to the IG-MHP without any data augmentation.

6.2 Instance-guided multi-human parsing method on non-overlapping human instances

The effectiveness of the proposed instance-guided multi-human parsing method on overlapping human instances has been confirmed by previous experiments, from both a qualitative and quantitative point of view. In this section, the instance-guided multi-human parsing approach is evaluated considering also images that do not include overlapping bodies. In order to do this, models presented so far have been tested on the whole CIHP validation set. Results are shown in Table 6.4. It is clear to see that, despite bringing improvements on CIHP subsets including occlusions between people, when considering also images without occlusions, the advantage gained is lost. The culprit of this is once again that fine-tuning the network on whole human instances ground-truth, leaving body part prediction almost unsupervised, will inevitably bring slight changes to predicted classes. Examples of this are shown in Figure 6.8, which illustrate how class prediction is affected by instance-guided fine-tuning. When facing challenging scenarios, such as large occluded people, the proposed approach is able to recover more accurate human instances, counter-balancing this deterioration in body part prediction. Figure 6.7, compares the evolution of the part-aware mean Intersection-over-

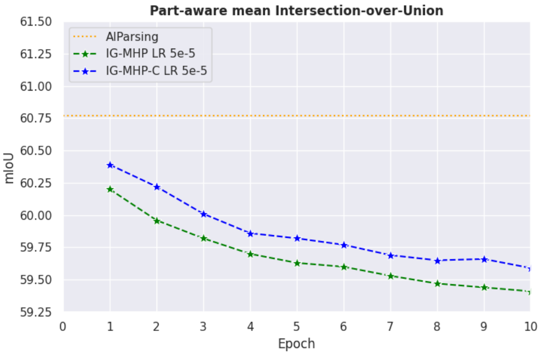


Figure 6.7: Evolution of the part-aware mean Intersection-over-Union on the full set of CIHP validation images for the best IG-MHP model found in the first experiment, without data augmentation, and for the IG-MHP model with identical learning rate and data augmentation strategy C.

	mIoU	pix_acc	mean_acc	mIoUh	AP^P_{50}	AP^P_{vol}	PCP
AIParsing	**60.77**	**90.29**	71.37	90.05	**76.01**	**60.47**	**69.22**
IG-MHP LR 3e-4	59.35	89.80	70.54	89.89	71.35	57.45	67.30
IG-MHP LR 5e-5	59.90	89.98	71.06	89.90	73.35	58.58	68.11
IG-MHP LR 3e-4 A	59.34	89.82	69.90	89.83	71.38	57.63	67.34
IG-MHP LR 3e-4 B	59.54	89.88	70.03	89.92	72.81	57.52	68.10
IG-MHP LR 3e-4 C	60.39	90.13	**71.44**	**90.15**	74.47	59.25	68.76

Table 6.4: Performances of the best models from the experiments in Section 6.1 on the whole CIHP validation set.

Union on the full set of CIHP validation images for the best IG-MHP model found in the first experiment, without data augmentation, and for the IG-MHP model with identical learning rate and data augmentation strategy C (image appearance, flipping and cropping).

To confirm this intuition, Table 6.5 shows evaluation metrics computed combining AIParsing baseline predictions with predictions from the proposed IG-MHP method. Specifically, on images having degree of overlap at least equal to 0.2, the new approach is applied. On images that do not frame overlapping people, AIParsing results are considered. It can be easily observed that, in this case, the advantage gained by the instance-guided approach is not cancelled out. This further proves that the degradation in performances that is found on the full set of CIHP images is due to images that frame people between which there is little or no overlapping, that is, cases that do not benefit from an enhancement in human segmentation quality.

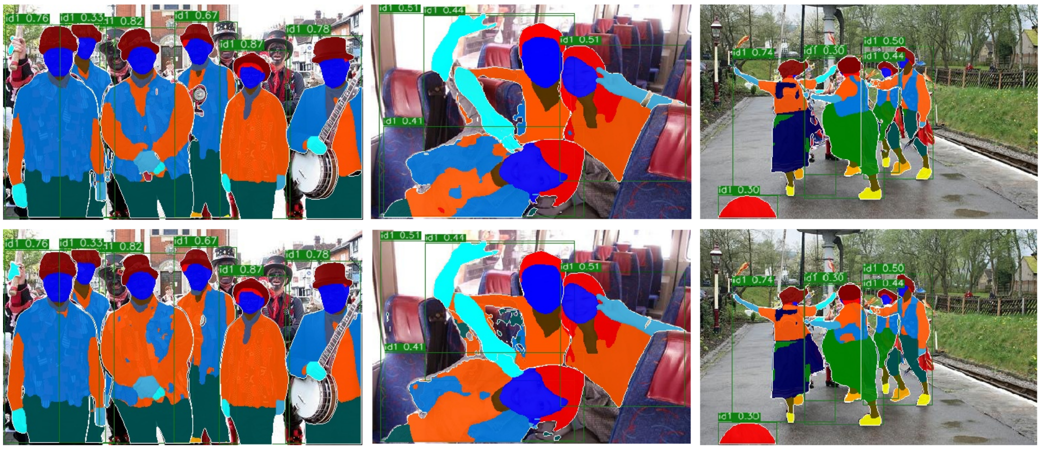


Figure 6.8: Examples of images affected by body part predictions errors after fine-tuning. Comparison between AIParsing multi-human parsing results (top row) and single-view method results (bottom row).

	mIoU	pix_acc	mean_acc	mIoUh	AP^P_{50}	AP^P_{vol}	PCP
AIParsing	60.77	90.29	71.37	90.05	76.01	60.47	69.22
IG-MHP LR 5e-5	60.94	90.32	71.98	90.40	75.88	60.44	69.78
IG-MHP LR 3e-4	60.71	90.24	71.77	90.39	74.59	59.89	69.30
IG-MHP-A	60.69	90.25	71.42	90.32	74.60	59.88	69.30
IG-MHP-B	60.79	90.27	71.49	90.37	74.76	59.97	69.36
IG-MHP-C	**61.14**	**90.39**	**72.13**	**90.42**	**76.06**	**60.56**	**69.91**

Table 6.5: Performances of the best models from the experiments in Section 6.1 on the whole CIHP validation set. AIParsing predictions are considered for images not including overlaps, while IG-MHP models predictions are considered for images including overlaps.

Chapter 7

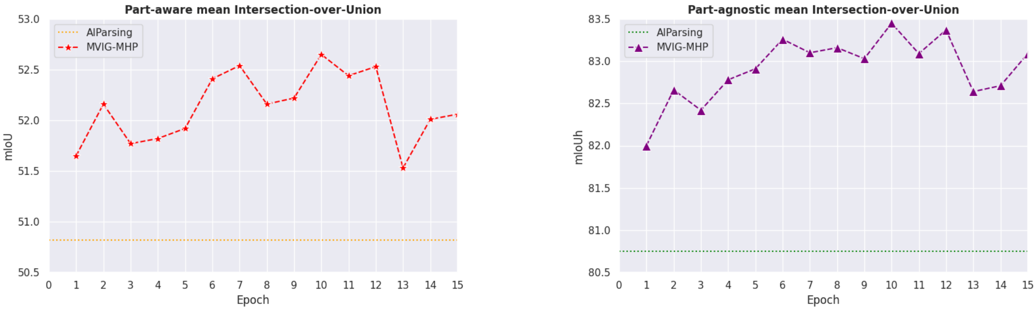
Experimental results for multi-view instance-guided multi-human parsing

In this chapter, the multi-view instance-guided multi-human parsing approach presented in Section 5.2 will be evaluated through different experiments, with the objective of improving the performance of a state-of-the-art architecture (i.e., AIParsing) in the challenging scenario of strong occlusions between people. With respect to the instance-guided multi-human parsing method validated in Chapter 6, the multi-view instance-guided approach makes use of a multi-view loss function to impose instance identity and body part prediction consistency across views. As done for the experimental validation presented in Chapter 6, the AIParsing architecture and its performance on the CIHP dataset will be considered as baseline. In particular, different subsets of images with increasing degree of overlap will be considered, in order to highlight the advantage of exploiting multi-view information to improve segmentation accuracy on images with strong occlusions between people. Furthermore, an ablation study on the role of different numbers of multi-view images for multi-view consistency loss computation will be carried out.

7.1 Multi-view instance-guided multi-human parsing on overlapping human instances

The proposed multi-view instance-guided multi-human parsing approach is now evaluated considering overlapping human instances, comparing obtained results with the AIParsing baseline. For this experiment, AIParsing has been fine-tuned on CMU Kinoptic-HIS dataset for 15 epochs, using a learning rate equal to 3e-4. Experiments for instance-guided multi-human parsing on occlusions, in fact, testify how the approach benefits from a higher learning rate. Multi-view loss consistency is computed taking into account views coming from two adjacent viewpoints and for each 3D human instance, the number of projections considered is 50, with maximum distance between each point and the point closest to the camera equal to 30 cm. The batch size is equal to 2, as the number of views used by the multi-view loss. As done for experiments described in Chapter 6, the best model is selected according to the value of the part-aware mean Intersection-over-Union computed CIHP_DoO_60. To mitigate the domain gap between fine-tuning and evaluation datasets, data augmentation strategy A, as defined in Section 6.1.2, is adopted. It combines background replacement, gaussian blur and jitter on illumination and contrast. Data augmentation affecting image geometry and orientation (e.g., image cropping and flipping) is not compatible with the computation of the multi-view loss and, thus, is not considered for these experiments. The trend for part-aware mean Intersection-over-Union during fine-tuning is illustrated in Figure 7.1a. The best model is obtained after 10 epochs. It achieves part-aware mIoU equal to 52.65 which corresponds to an improvement of the 3.6% with the respect to the AIParsing baseline. It also tops the mIoU reached by IG-MHP (i.e., instance-guided multi-human parsing with single-view losses) using the same learning rate (52.57), even though the two models are not directly comparable as they use different batch sizes (2 vs. 8 images).

Table 7.1 shows overall model performances on CIHP_DoO_40, CIHP_DoO_60 and CIHP_DoO_80 subsets. MVIG-MHP tops the baseline in terms of part-aware mIoU on subsets with degree of overlap equal to 0.6 and 0.8, which confirms an enhancement in multi-human parsing quality in case of strong occlusions. This appears to be due to improved human instance segmentation, as it can be observed that the part-agnostic mean Intersection-over-Union (mIoUh) gains a significant advantage with respect to the AIParsing baseline considering all three



(a) Part-aware mean Intersection-over-Union for the MVIG-MHP method.

(b) Part-agnostic mean Intersection-over-Union for the MVIG-MHP method.

Figure 7.1: Part-aware mean Intersection-over-Union and part-agnostic mean Intersection-over-Union for the multi-view instance-guided multi-human parsing (MVIG-MHP) method.

	Model	mIoU	pix_acc	mean_acc	mIoUh	AP_{50}^{P}	AP_{vol}^{P}	PCP
DoO_40	AIParsing	**55.42**	87.57	**64.95**	86.52	**68.75**	**56.01**	**63.24**
	MVIG-MHP	55.13	**87.58**	64.69	**87.61**	62.74	52.78	60.79
DoO_60	AIParsing	50.82	84.84	59.47	80.75	**61.09**	**51.56**	**57.72**
	MVIG-MHP	**52.65**	**85.57**	**61.92**	**83.45**	55.85	48.64	54.96
DoO_80	AIParsing	46.81	83.08	58.25	79.21	**58.43**	**50.36**	**55.59**
	MVIG-MHP	**47.30**	**84.33**	**58.72**	**82.68**	55.42	47.52	54.29

Table 7.1: Overall performances of multi-view instance-level human parsing on CIHP_DoO_40, CIHP_DoO_60 and CIHP_DoO_80 subsets.

subsets (e.g., +4.38% on CIHP_DoO_80). This is consistent to what is illustrated in Figure 7.1b, showing a significant growth in human segmentation accuracy during fine-tuning. This highlights the effectiveness of the multi-view consistency in leading to more accurate and fuller human masks.

Also from a quantitative point of view, the advantages of the multi-view approach are remarkable. Compared to the baseline AIParsing and the single-view loss approach, by exploiting multi-view information, the network can better distinguish human instances and recover any holes in the instances due to non-predicted or non-associated body parts. This is evident from Figure 7.2, showing a comparison between results given by the AIParsing baseline, IG-MHP and MVIG-MHP, proposed in this thesis. The boost provided by multi-view consistency in human instance segmentation is strong. MVIG-MHP is able to segment even the fifth person behind all the others, ignored by the other methods.

Discrimination effects on overlapped human instances are highlighted in Figure 7.3. Considering occluded people on the left side of the image, a correct multi-human parsing result should include predictions about one of the two instances only, as only one bounding box was produced by the detection head of the baseline top-down architecture AIParsing. AIParsing results, however, clearly

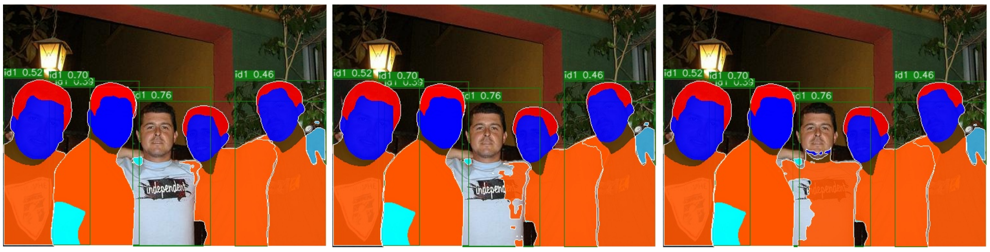


Figure 7.2: Comparison between results from AIParsing (first image), Instance-guided multi-human parsing (second image) and Multi-view instance-guided multi-human parsing (third image), shown to highlight network human instance recovery capabilities.

show some issues related to foreground instance discrimination, as parts belonging to both people appear in the output parsing map. While instance-guided multi-human parsing fails at improving disambiguation, the additional multi-view loss exploited by MVIG-MHP results in improved instance discrimination. MVIG-MHP chooses the woman as target instance and only few portions of the close man body (right hand and part of face) appear to be mismatched.

Considering body part prediction, unfortunately, multi-view consistency does not seem to produce the desired results. As depicted in Figure 7.4, MVIG-MHP classifies part of the skirt of the girl in foreground as dress. This effect is confirmed by the instance-level metrics reported in Table 7.1 on all three CIHP overlap subsets considered, which obtain lower values than the baseline network.

Table 7.2 show performances on CIHP_DoO_60 for the AIParsing baseline, instance-guided multi-human parsing and multi-view instance-guided multi-human parsing best models. IG-MHP and MVIG-MHP models compared have been trained adopting the same learning rate and data augmentation strategy. The multi-view consistency loss leads to improving both human parsing and human instance segmentation performances with respect to the AIParsing baseline and to IG-MHP, employing single-view loss functions only. As already pointed out, however, multi-view consistency leads to decreased accuracy in body part prediction with respect to instance-guided multi-human parsing with single-view losses, as shown by instance-level metrics. This means that the advantage obtained by MVIG-MHP on part-aware mean Intersection-over-Union is due to improved human instance discrimination and recovery, with respect to instance-guided multi-human parsing, as shown in Figures 7.3 and 7.2.

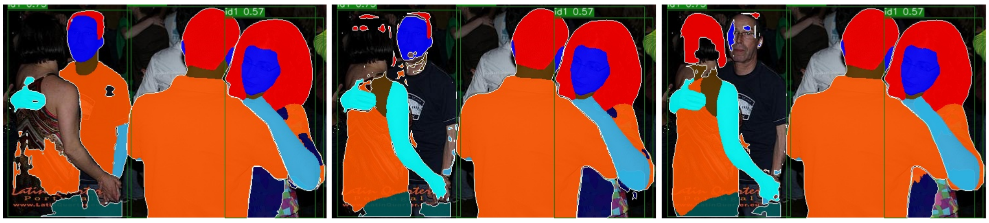


Figure 7.3: Comparison between results from AIParsing (first image), Instance-guided multi-human parsing (second image) and Multi-view instance-guided multi-human parsing (third image), shown to highlight network human instance discrimination capabilities.

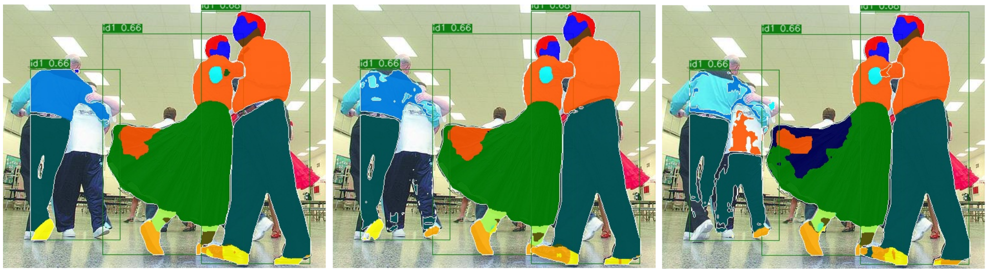


Figure 7.4: Comparison between results from AIParsing (first image), Instance-guided multi-human parsing (second image) and Multi-view instance-guided multi-human parsing (third image), shown to highlight effects on body parts prediction accuracy.

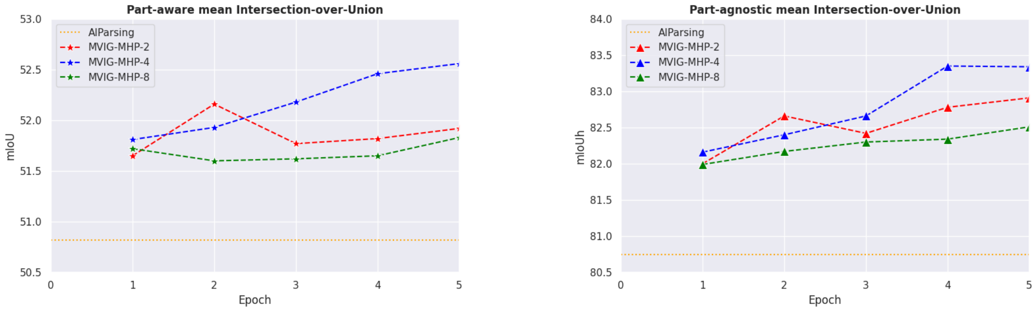
7.1.1 Analysis on number of views

With the objective on investigating and analysing the role of different numbers of views for multi-view consistency constraints, a series of experiments considering a varying number of views is carried out. In particular, imposing consistency taking into account contributions given by a larger number of views could provide more reliable results. At the same time, however, if the different viewpoints are far from each other, exploiting contributions coming from many views could introduce noise. Consequently, imposing consistency using noisy and unreliable labels, could be detrimental to overall performance.

Experiments are carried out considering sets of 2, 4 and 8 adjacent views, using image batch sizes equal to 2, 4 and 8, respectively. Figure 7.5a and 7.5b show the evolution of part-aware mIoU and part-agnostic mIoU over the fine-tuning process for each conducted experiment. Table 7.3 shows the results on

Model	mIoU	pix_acc	mean_acc	mIoUh	AP^P_{50}	AP^P_{vol}	PCP
AIParsing	50.82	84.84	59.47	80.75	**61.09**	**51.56**	**57.72**
IG-MHP	52.47	85.51	61.81	83.00	56.99	49.19	55.94
MVIG-MHP	**52.65**	**85.57**	**61.92**	**83.45**	55.85	48.64	54.96

Table 7.2: Overall performances on CIHP_DoO_60 for the AIParsing baseline, instance-guided multi-human parsing and multi-view instance-guided multi-human parsing best models.



(a) Part-aware mean Intersection-over-Union for MVIG-MHP method.

(b) Part-agnostic mean Intersection-over-Union for MVIG-MHP method.

Figure 7.5: Part-aware and part-agnostic mean Intersection-over-Union for multi-view instance-guided multi-human parsing method.

Model	N. Views	mIoU	pix_acc	mean_acc	mIoUh	AP_{50}^{P}	AP_{vol}^{P}	PCP
AIParsing	-	50.82	84.84	59.47	80.75	**61.09**	**51.56**	**57.72**
MVIG-MHP	2	52.16	85.41	61.41	82.66	56.50	49.44	56.19
MVIG-MHP	4	**52.56**	**85.71**	**62.10**	**83.34**	57.04	49.56	56.41
MVIG-MHP	8	51.83	85.34	61.07	82.51	56.81	49.55	56.39

Table 7.3: Overall performances of MVIG-MHP models using 2, 4 and 8 multi-view images for multi-view consistency loss computation, on CIHP_DoO_60.

CIHP_DoO_60 obtained by the best models. It can be observed that the model achieving the best performances in terms of part-aware mean Intersection-over-Union is MVIG-MHP using a multi-view consistency loss on 4 multi-view images, with batch size equal to 4.

It is interesting to notice that the model exploiting the highest number of views is the one that performs the worst in comparison to the others. The reason for this could be that, considering 8 adjacent views, the multi-view loss term ends up fusing contributions coming from viewpoints that are very far one from the other, aggregating many predictions that do not agree. Figure 7.6 shows a sequence of 8 adjacent views from CMU Kinoptic-HIS, parsed by AIParsing baseline. The black dots in each view represents the forward projections of a same 3D point. It is clear to see that, in such case, multi-view consistency would fuse many disagreeing contributions and would not provide useful supervision for body part prediction.

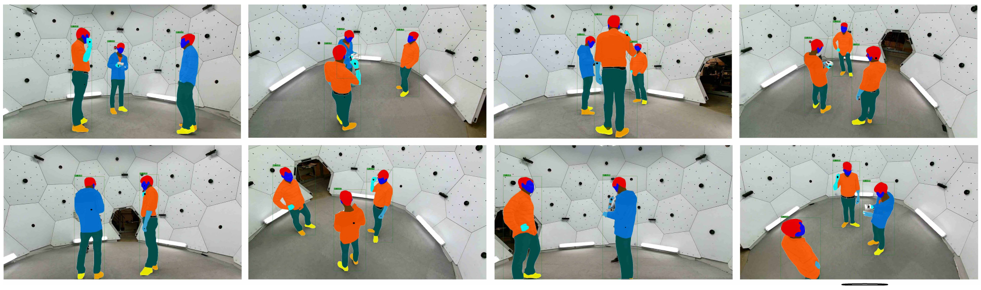


Figure 7.6: Sequence of 8 adjacent views from CMU Kinoptic-HIS, parsed by AIParsing baseline. The black dots in each view represents the forward projections of a same 3D point.

7.2 Multi-view instance-guided multi-human parsing on non-overlapping human instances

The effectiveness of the proposed multi-view instance-guided multi-human parsing method on overlapping human instances has been validated, from both a qualitative and quantitative point of view, by experiments presented in the previous section. Here, the proposed approach is evaluated considering also images that do not include occlusions between bodies. To do this, models presented so far have been tested on the whole CIHP validation set. Results are shown in Table 7.4. The results obtained exploiting multi-view consistency are compared with those obtained by the instance-guided multi-human parsing approach (IG-MHP). In particular, among the models presented in Chapter 6, the ones which obtained the best results on non-overlapped human instances, are considered for the comparison. These include IG-MHP trained using a learning rate equal to 5e-5, without any data augmentation, and IG-MHP-C, trained using a learning rate equal to 3e-4, using data augmentation strategy C, as defined in Section 6.1.. These are compared with MVIG-MHP models achieving the best overall performances on scenarios with overlapped human instances. These are MVIG-MHP-2 from the experiment presented in Section 7.1 and MVIG-MHP-4, presented in the previous experiment. Results on CIHP validation set are a confirmation of what was already found analysing previous experiments. Examining both global- and instance-level metrics, performances of the models exploiting multi-view consistency testify a loss in body part prediction accuracy with respect to the AIParsing baseline and to the single-view loss instance-guided MHP models. Comparing multi-view instance-guided models, multi-view consistency exploiting 4 views is the most effective. Consistently to what was shown for instance-guided multi-human parsing in Chapter 6, despite improving on CIHP

	mIoU	pix_acc	mean_acc	mIoUh	AP_{50}^{P}	AP_{vol}^{P}	PCP
AIParsing	**60.77**	**90.29**	71.37	90.05	**76.01**	**60.47**	**69.22**
IG-MHP LR 5e-5	59.90	89.98	71.06	89.90	73.35	58.58	68.11
IG-MHP-C LR 3e-4	60.39	90.13	**71.44**	**90.15**	74.47	59.25	68.76
MVIG-MHP-2	59.12	89.72	69.49	89.91	70.54	57.03	66.69
MVIG-MHP-4	59.67	89.93	70.55	89.98	71.99	57.85	67.59

Table 7.4: Performances on the full CIHP validation set of the AIParsing baseline, the best IG-MHP models and the best MVIG-MHP models.

	mIoU	pix_acc	mean_acc	mIoUh	AP_{50}^{P}	AP_{vol}^{P}	PCP
AIParsing	60.77	90.29	71.37	90.05	76.01	60.47	69.22
IG-MHP LR 5e-5	60.94	90.32	71.98	90.40	75.88	60.44	69.78
IG-MHP-C LR 3e-4	**61.14**	**90.39**	**72.13**	**90.42**	**76.06**	**60.56**	**69.91**
MVIG-MHP-2	60.60	90.19	71.23	90.34	74.37	59.69	69.08
MVIG-MHP-4	60.71	90.27	71.76	90.41	75.17	60.08	69.50

Table 7.5: Performances on the full CIHP validation set of the AIParsing baseline, the best IG-MHP models and the best MVIG-MHP models, considering AIParsing results for images without occlusions and the proposed method for images with overlapped instances.

subsets including occlusions between people, when considering also images without occlusions, the advantage gained by the proposed method is lost because of body part prediction degradation. In particular, models exploiting single-view loss functions only hold better performances as body part prediction is better preserved.

To better evaluate the approach contribution, Table 7.5 show the results on the full CIHP validation set, considering AIParsing baseline on images without occlusions and the proposed methods IG-MHP and MVIG-MHP on images from CIHP_DoO_20. It can be observed how instance-guided multi-human parsing models still perform better than models including multi-view consistency, as constraints on body part predictions coherence across views do not provide the desired results.

Chapter 8

Conclusions

This thesis addressed multi-human parsing in the challenging scenario of severe occlusions in the image. Despite producing significant results on multi-human parsing datasets, current deep-learning-based approaches still struggle to properly deal with overlapping human instances. An in-depth analysis on state-of-the-art methods, taking into consideration images with different occlusion severity, highlights important issues in accurately parsing human bodies if largely overlapped. When human bodies appear to be extensively overlapped, respective body parts intersect. This causes current multi-human parsing models to struggle in associating body parts to the correct human instance, producing incomplete and inaccurate parsing maps. Exploiting 3D information about humans in the scene resolves ambiguity between instances, as people appear as separated if framed from a different point of view. Following this intuition, this thesis proposes to take advantage of multi-view information in order to enhance multi-human parsing on strong occlusions between bodies.

A novel learning framework is proposed to improve multi-human parsing performance, exploiting human instance segmentation ground-truth as auxiliary information to guide network learning. Learning on the new data is guided by single-view human instance segmentation losses, aiming at improving foreground human instance discrimination, as well as human instance segmentation quality and outline. Novel multi-view consistency loss terms are used to enforce coherent instance and body part predictions across multiple views of the same scene, to enhance overlapping human instances discrimination and provide sparse supervision to body part segmentation. A state-of-the-art multi-human parsing architecture is fine-tuned exploiting the weak supervision provided by multi-view human instance segmentation ground-truth, in order to enhance human instance parsing on strong occlusions.

The proposed learning framework has been validated considering the AIParsing architecture, state-of-the-art model for multi-human parsing, and the CIHP dataset. Experimental results show the effectiveness of the approach that, on images presenting extensive occlusions between human bodies, is able to significantly improve body parts and human instance segmentation quality, achieving an advantage in terms of part-aware mean Intersection-over-Union up to the 4.25% with respect to the AIParsing network. The main criticality remains the one related to the representation gap between multi-human parsing datasets, including both human instance discrimination and fine-grained body parts labels, and human instance segmentation ground-truth used to guide the fine-tuning, providing coarser granularity information. This leads to a degradation in body part prediction accuracy that, when facing very challenging scenarios, such as largely occluded people, is counter-balanced by the amount of human body that the model is now able to recover. When evaluated on a broader set of images in which occlusions are not so critical, such as the entire CIHP validation set, this effect becomes visible and the gained advantage is very diminished. The contribution of a body parts prediction consistency constraint across multiple views, is not sufficient to completely solve the issue.

To conclude, the objective of this thesis was indeed satisfactorily achieved, as the proposed approach is able to improve multi-human parsing performance on the very challenging scenario in which people are strongly occluded one with the other. Future developments and research will be devoted to mitigating the detrimental effect found on body parts segmentation accuracy, which would allow to enhance multi-human parsing also on images which do not present occlusions. As the proposed approach does not depend on a particular multi-human parsing architecture or on a predefined set of body parts of interest, the generalization capabilities of the proposed approach will also be evaluated, applying the devised learning framework to other networks and datasets, such as LV-MHP-v2.0, particularly challenging as providing annotations for more than 50 semantic classes.

References

[1] Wenguan Wang, Yuanlu Xu, Jianbing Shen, and Song-Chun Zhu. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4271–4280, 2018.

[2] Haoye Dong, Xiaodan Liang, Yixuan Zhang, Xujie Zhang, Xiaohui Shen, Zhenyu Xie, Bowen Wu, and Jian Yin. Fashion editing with adversarial parsing learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8120–8128, 2020.

[3] Matteo Terreran, Leonardo Barcellona, Daniele Evangelista, and Stefano Ghidoni. Multi-view human parsing for human-robot collaboration. In *2021 20th International Conference on Advanced Robotics (ICAR)*, pages 905–912. IEEE, 2021.

[4] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018.

[5] Yang Wang, Duan Tran, Zicheng Liao, and David Forsyth. Discriminative hierarchical part-based models for human parsing and action recognition. *Gesture Recognition*, pages 273–301, 2017.

[6] Chia-Wei Hsieh, Chieh-Yun Chen, Chien-Lung Chou, Hong-Han Shuai, Jiaying Liu, and Wen-Huang Cheng. Fashionon: Semantic-guided image-based virtual try-on with detailed human and clothing information. In *Proceedings of the 27th ACM international conference on multimedia*, pages 275–283, 2019.

[7] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards multi-pose guided virtual try-on network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9026–9035, 2019.

[8] Tianfei Zhou, Wenguan Wang, Si Liu, Yi Yang, and Luc Van Gool. Differentiable multi-granularity human representation learning for instance-aware human seman-

tic parsing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1622–1631, 2021.

[9] Zhiwei Liu, Xiangyu Zhu, Lu Yang, Xiang Yan, Ming Tang, Zhen Lei, Guibo Zhu, Xuetao Feng, Yan Wang, and Jinqiao Wang. Multi-initialization optimization network for accurate 3d human pose and shape estimation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1976–1984, 2021.

[10] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1062–1071, 2018.

[11] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 346–363. Springer, 2020.

[12] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1971–1978, 2014.

[13] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 770–785, 2018.

[14] Jian Zhao, Jianshu Li, Yu Cheng, Terence Sim, Shuicheng Yan, and Jiashi Feng. Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 792–800, 2018.

[15] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[16] Lu Yang, Qing Song, Zhihui Wang, and Ming Jiang. Parsing r-cnn for instance-level human analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 364–373, 2019.

[17] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2018.

[18] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 542–551, 2019.

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

[20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[21] 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.

[22] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[24] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[25] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018.

[26] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9157–9166, 2019.

[27] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019.

[28] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[29] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[30] Sanyi Zhang, Xiaochun Cao, Guo-Jun Qi, Zhanjie Song, and Jie Zhou. Aiparsing: anchor-free instance-level human parsing. *IEEE Transactions on Image Processing*, 31:5599–5612, 2022.

[31] Tao Ruan, Ting Liu, Zilong Huang, Yunchao Wei, Shikui Wei, and Yao Zhao. Devil in the details: Towards accurate single and multiple human parsing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 4814–4821, 2019.

[32] Haoyu He, Jing Zhang, Qiming Zhang, and Dacheng Tao. Grapy-ml: Graph pyramid mutual learning for cross-dataset human parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10949–10956, 2020.

[33] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3260–3271, 2020.

[34] Lu Yang, Wenhe Jia, Shan Li, and Qing Song. Deep learning technique for human parsing: A survey and outlook. *arXiv preprint arXiv:2301.00394*, 2023.

[35] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[36] Sanyi Zhang, Guo-Jun Qi, Xiaochun Cao, Zhanjie Song, and Jie Zhou. Human parsing with pyramidical gather-excite context. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(3):1016–1030, 2020.

[37] Wenguan Wang, Hailong Zhu, Jifeng Dai, Yanwei Pang, Jianbing Shen, and Ling Shao. Hierarchical human parsing with typed part-relation reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8929–8939, 2020.

[38] Lu Yang, Qing Song, Zhihui Wang, Mengjie Hu, Chun Liu, Xueshi Xin, Wenhe Jia, and Songcen Xu. Renovating parsing r-cnn for accurate multiple human parsing. In *European Conference on Computer Vision*, pages 421–437. Springer, 2020.

[39] 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*, 40(4):834–848, 2017.

[40] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[41] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[42] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 516–520, 2016.

[43] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4413–4421, 2018.

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

[45] Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):267–282, 2007.

[46] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *Proceedings of 28th British Machine Vision Conference*, pages 1–13, 2017.

[47] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6399–6408, 2019.

[48] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 196–214. Springer, 2020.

[49] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

[50] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[51] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[52] Akin Caliskan, Armin Mustafa, Evren Imre, and Adrian Hilton. Multi-view consistency loss for improved single-image 3d reconstruction of clothed people. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[53] Morris Antonello, Daniel Wolf, Johann Prankl, Stefano Ghidoni, Emanuele Menegatti, and Markus Vincze. Multi-view 3d entangled forest for semantic segmentation and mapping. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1855–1862. IEEE, 2018.