Differentiable Multi-Granularity Human Representation Learning for Instance-Aware Human Semantic Parsing

Author(s):
Zhou, Tianfei; Wang, Wenguan; Liu, Si; Yang, Yi; Van Gool, Luc

Publication date:
2021

Permanent link:
https://doi.org/10.3929/ethz-b-000519398

Originally published in:
https://doi.org/10.1109/CVPR46437.2021.00167

Differentiable Multi-Granularity Human Representation Learning for Instance-Aware Human Semantic Parsing

Tianfei Zhou[1] , Wenguan Wang[1*] , Si Liu[2] , Yi Yang[3] , Luc Van Gool[1]
[1]Computer Vision Lab, ETH Zurich [2]Institute of Artificial Intelligence, Beihang University [3]University of Technology Sydney

https://github.com/tfzhou/MG-HumanParsing

Abstract

To address the challenging task of instance-aware human part parsing, a new bottom-up regime is proposed to learn category-level human semantic segmentation as well as multi-person pose estimation in a joint and end-to-end manner. It is a compact, efficient and powerful framework that exploits structural information over different human granularities and eases the difficulty of person partitioning. Specifically, a dense-to-sparse projection field, which allows explicitly associating dense human semantics with sparse keypoints, is learnt and progressively improved over the network feature pyramid for robustness. Then, the difficult pixel grouping problem is cast as an easier, multi-person joint assembling task. By formulating joint association as maximum-weight bipartite matching, a differentiable solution is developed to exploit projected gradient descent and Dykstra's cyclic projection algorithm. This makes our method end-to-end trainable and allows back-propagating the grouping error to directly supervise multi-granularity human representation learning. This is distinguished from current bottom-up human parsers or pose estimators which require sophisticated post-processing or heuristic greedy algorithms. Experiments on three instance-aware human parsing datasets show that our model outperforms other bottom-up alternatives with much more efficient inference.

1. Introduction

Instance-aware human parsing, *i.e.*, partitioning humans into semantic parts (*e.g.*, torso, head) and associating each part with the corresponding human instance, has only started to be tackled in the literature (dating back to [31]). This article addresses this task through a new regime, which learns to jointly estimate human poses (*i.e.*, a sparse, skeleton-based human representation) and segment human parts (*i.e.*, a pixel-wise, fine-grained human representation) in an *end-to-end trainable*, *bottom-up* fashion.

*Corresponding author: *Wenguan Wang*.

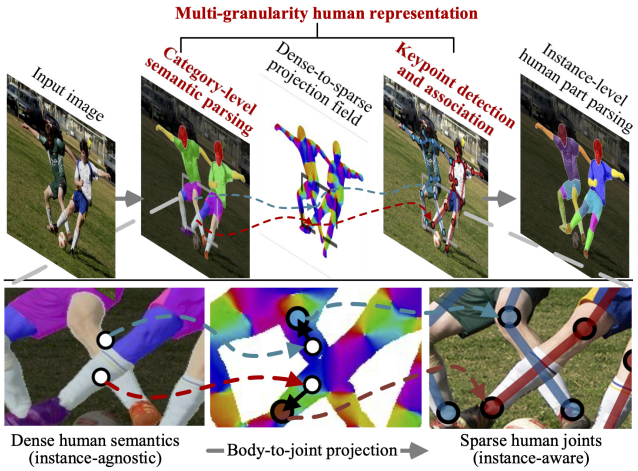


Figure 1: **Overview of our new bottom-up regime for instance-aware human semantic parsing.** By learning **1)** *category-level* human semantic parsing, **2)** *body-to-joint* projection, and **3)** *bottom-up* keypoint detection and association in a joint and end-to-end manner, our model tackles the task in a differentiable, multi-granularity human representation learning framework.

In the field of human parsing, the idea of leveraging human pose as structural knowledge to facilitate human understanding has been exploited for years [56]. However, previous efforts only focus on the *instance-agnostic* setting [54, 53, 15, 37, 64, 48]. Further, most of them directly utilize human joints (pre-detected from off-the-shelf pose estimators) as extra inputs [56, 54], or simply generate keypoint estimations as a by-product [15, 64]. In sharp contrast, by learning to associate semantic person pixels with their closest person instance joints, our model seamlessly injects bottom-up pose estimation into instance-aware human semantic learning and inference. Thus, our human parser can make use of the complementary cues from sparse human joints and dense part semantics, and push further the envelope of human understanding in unconstrained environments. This represents an early effort to formulate instance-aware human parsing and multi-person pose estimation in a bottom-up, differentiable, and multi-granularity human representation learning framework (see Fig. 1).

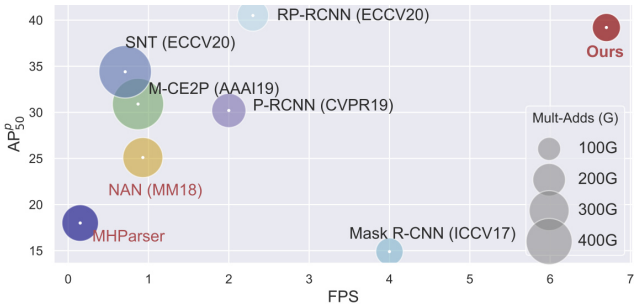


Figure 2: **Trade-off between performance *vs.* efficiency** on MHP_{v2} `val` [65]. The x-axis and y-axis denote FPS and AP^p_{50}, respectively. The circle size indicates Multi-Adds (G). Top-down and bottom-up models are given in black and red, respectively. Our model shows promising performance with high efficiency.

More importantly, our framework yields a new paradigm for bottom-up, instance-aware human part parsing. For human instance discrimination, current bottom-up human parsers learn to associate pixels by directly regressing instance locations [65] or predicting pair-wise connectiveness between pixels [30]. However, instance locations (*i.e.*, human centroids or bounding box corners) are less semantic and pair-wise embeddings are difficult to learn (due to deformations, variations and occlusions of human bodies) [50]. As shown in Fig. 1, our model instead regresses semantic- and geometry-rich human joints as pixel embeddings. Then, fine-grained human semantics can be efficiently grouped through body joint association. In essence, it formulates instance-aware human semantic parsing by jointly learning: **1)** *category-level* human semantic parsing, **2)** *body-to-joint* projection, and **3)** *bottom-up* multi-person keypoint detection and association. Thus, our model avoids sophisticated inference and heavy network designs, and neatly explores the structural constraints over human bodies. Such a flexible network design can benefit from progress in bottom-up pose estimation and semantic segmentation techniques, and significantly differentiates itself from existing instance-aware human parsers.

Concretely, three crucial techniques, for the first time in the field, are exploited to deliver our compact and powerful instance-aware human parsing solution:

- *Differentiable body-to-joint projection*: We propose a Dense-to-Sparse Projection Field (DSPF), which is a set of 2D vectors over the image lattice. For each human pixel, its DSPF vector encodes the offset to the closest instance joint. DSPF thus allows us to explicitly associate dense human semantics with sparse joint representations.
- *Multi-step DSPF estimation*: To address the difficulties of human body variations and occlusions, DSPF is computed in a coarse-to-fine fashion, over the network feature pyramid. Deeper-layer features are low-resolution, yet robust to the challenges. They are thus used to derive an initial discriminative DSPF. Conditioning on the shallower-layer features, a finer DSPF is inferred by computing the residue to the coarse estimation. To reduce the feature-space distance across network layers, cross-layer feature alignment is learnt and adopted. In this way, the residue has a smaller magnitude and is easier to infer.
- *An end-to-end trainable framework*: Current bottom-up multi-person pose estimators solve joint association through heuristic greedy algorithms [4, 27], independent from joint detection. This breaks the end-to-end pipeline and leads to suboptimal results. Since joint association can be formulated as maximum-weight bipartite matching [4], we explore a differentiable solution to this, inspired by [61]. It revisits projected gradient descent and Dykstra's cyclic projection [2] for convex constrained minimization. The solution is neat and light-weight, allowing directly using the grouping errors for supervision.

Our model is also distinguished for its practical utility. First, it can benefit a host of human-centric applications. Some of them, such as augmented reality, are pose guided, while others, such as live streaming, video editing and virtual try-on systems, require understanding of fine-grained human semantics. Thus, our method can meet different needs in real-life applications with only a *single* model. Second, due to its bottom-up nature and fast keypoint association, our method generates instance-level parsing results at a high speed of 0.15 s per image (see Fig. 2), irrespective of the number of people in the scene, and, which is much faster than other alternatives (*e.g.*, 1.15 s [42], 14.94 s [30]).

Extensive experiments are conducted on three instance-aware human parsing datasets. Specifically, on MHP_{v2} [65], our model achieves an AP^p_{50} of 39.0, much better than current top-leading bottom-up method (*i.e.*, NAN [65] of 25.1) and on par with best top-down parsers [24, 57]. On DensePose-COCO [1] and PASCAL-Person-Part [53], our model also produces compelling performance. Overall, our model shows favorable results with a fast speed.

2. Related Work

Instance-Aware Human Semantic Parsing. Fine-grained human semantic segmentation, as one of the central tasks in human understanding, has applications in human-centric vision [41, 66, 43], human-robot interaction [11] and fashion analysis [46]. However, previous studies mainly focus on *category-level* human parsing [32, 15, 12, 35, 47, 63, 64, 49]; only very few human parsers are specifically designed for the *instance-aware* setting. As of to date, there exist two paradigms for instance-aware human parsing: *top-down* and *bottom-up*. Top-down approaches [31, 58, 42, 24, 57] typically locate human instance proposals first, and then parse each proposal in a fine-grained manner. In contrast, bottom-up human parsers [14, 30, 65] simultaneously perform pixel-wise instance-agnostic parsing and pixel grouping, inspired by existing bottom-up instance segmentation techniques. Their grouping strategies vary from instance-

edge-aware clustering [14], to graph-based superpixel association [30], to proposal-free instance localization [65].

Our model falls into the bottom-up paradigm and has several unique characteristics. First, our method formulates instance-aware human part parsing as a multi-granularity human representation learning task. However, other alternatives, whether top-down or bottom-up, fail to explore skeleton-based human structures. Second, our model yields a new bottom-up regime for instance-aware human part parsing. Through projecting dense human semantics into sparse body joints, the task of pixel grouping can be easily tackled by joint association. Third, our method achieves promising results with much improved efficiency, leading to high utility in downstream applications.

Multi-Person Pose Estimation. Allocating body joints to human instances, as another representative human understanding task, has been extensively studied over the past decades. Also, current popular solutions for multi-person pose estimation can be roughly categorized as either *top-down* or *bottom-up* methods. Specifically, top-down methods [13, 19, 40, 7] conduct single-person pose estimation over each pre-detected human bounding-box. Contrarily, bottom-up methods are detection-free; they usually predict identity-free keypoints, which are then assembled into valid pose instances [22, 4, 36, 27], or directly regress poses from person instance centroids in a single-stage fashion [38].

In our method, human joints are set as the targets for person pixel embedding learning, simplifying the procedure of pixel grouping. This flexible architecture can be seamlessly integrated with any bottom-up pose estimators in principle, and explicitly encodes human structural constraints. Hence, we formulate joint association as a *differentiable* matching problem [61], rather than relying on sophisticated post-processing [4, 27] like conventional bottom-up methods. Though [25] also addresses joint association in an end-to-end manner, it needs to learn a complicated and heavy graph network and cannot guarantee optimality.

Human/Object Representation. From a slightly broader perspective, the way of representing visual elements is crucial for visual understanding. From classic rectangular boxes and masks, to recent point-based object modeling forms (*e.g.*, sparse points [59], contours [55, 50], grid masks [6], and dense points [60]), the community is continually pursuing a proper object representation for more effective processing, further advancing the development of object detection and segmentation. Point-based representations are thought to be promising, as both geometric and semantic cues are encoded. This insight is shared by several skeleton-based human understanding models. Despite the efforts made prior to the renaissance of deep learning [9, 54], some recent studies explore body poses as an extra representation granularity in category-level human parsing [53, 37] or instance-aware human full-body segmentation [39, 62]. However, [53, 37, 62] only utilize human poses as a shape prior for feature enhancement, instead of as an informative clue for person separation. Thus, their main ideas are far different from ours. Although [39] also views human joints as pixel embedding targets, our method **1)** focuses on fine-grained human semantic part parsing; **2)** coarse-to-fine estimates the projection from human dense semantics to sparse joints through residual learning and feature alignment; and **3)** formulates joint association in a differentiable manner, yielding an end-to-end trainable model.

3. Our Approach

As shown in Fig. 3, our model learns to: **1)** predict category-level human parts (§3.1), **2)** build a Dense-to-Sparse Projection Field(DSPF,§3.2), and **3)** conduct human joint detection and association (§3.3). With **1)**, we have fine-grained human semantics, but they are identification-irrelevant. With **2)**, we can explicitly associate dense human semantics with sparse joints, as DSPF encodes the displacements from each person pixel to the closet human joints. Instance-aware parsing can then be achieved by bottom-up joint association. At the same time, human structural cues can be embedded. Finally, with **3)**, joint association is achieved by a differentiable solver, allowing the error signals for grouping to be back-propagated to supervise keypoint detection and feature learning. In this way, we provide an instance-aware human parsing framework based on multi-granularity human representation learning, which works in a bottom-up manner and is end-to-end trainable.

3.1. Instance-Agnostic Human Part Parsing

Given an input image $I \in \mathbb{R}^{W\times H\times 3}$, a backbone network is first employed to extract an L-level feature pyramid, *i.e.*, $\{\boldsymbol{X}_l \in \mathbb{R}^{W_l\times H_l\times C_l}\}_{l=1}^{L}$, where $W_l = W/2^{l+1}$ and $H_l = H/2^{l+1}$. The feature pyramid comprehensively encodes multi-scale visual features, from the highest spatial resolution ($l=1$) to the lowest ($l=L$).

Given $\{\boldsymbol{X}_l\}_{l=1}^{L}$, a segmentation head $\mathcal{F}^{\text{Seg}}$ is applied to parse category-level human part semantics (see Fig. 3):

$$\boldsymbol{S} = \mathcal{F}^{\text{Seg}}(\{\boldsymbol{X}_l\}_{l=1}^{L}) \in [0,1]^{W_1\times H_1\times P}. \tag{1}$$

Here, P indicates the number of human semantic part categories. $\mathcal{F}^{\text{Seg}}$ is implemented as a decoder architecture [5], which makes full use of the multi-scale features $\{\boldsymbol{X}_l\}_{l=1}^{L}$ and estimates $\boldsymbol{S}$ with the highest spatial resolution.

3.2. Multi-Step Dense-to-Sparse Projection Field (DSPF) Estimation

Given $\{\boldsymbol{X}_l\}_{l=1}^{L}$, our network learns DSPF, *i.e.*, $\boldsymbol{D} \in \mathbb{R}^{W_1\times H_1\times 2}$, to associate dense human semantics with sparse keypoints (Fig. 3). Specifically, for each point $\boldsymbol{u} = (u, v)$ in an image lattice $\Omega \in \mathbb{R}^2$, its DSPF is a 2D vector $\boldsymbol{D}(\boldsymbol{u}) = (\Delta u, \Delta v)$, where $\boldsymbol{u} + \boldsymbol{D}(\boldsymbol{u}) = (u + \Delta u, v + \Delta v)$

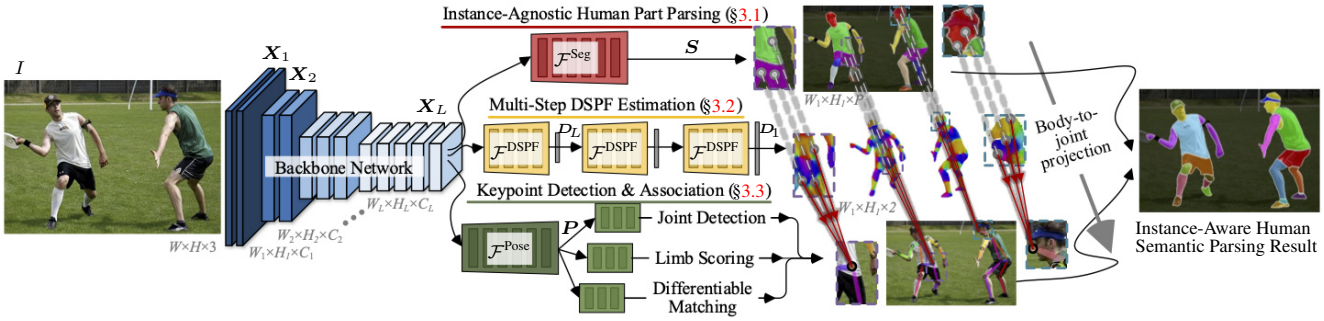


Figure 3: **Illustration of our multi-granularity human representation learning framework** for instance-aware human semantic parsing.

is expected to be the location of its nearest human instance joint. Note that $\boldsymbol{u}$ and $\boldsymbol{u}+\boldsymbol{D}(\boldsymbol{u})$ belong to the same human instance. Due to variations and deformations typically exhibited by human bodies, as well as ambiguities caused by close or occluded instances, it is difficult to directly infer DSPF. Thus, we estimate it in a coarse-to-fine framework.

Coarse-to-Fine Estimation. Given the coarsest feature (*i.e.*, $\boldsymbol{X}_L$), which has the largest receptive field and thus is good at robust, long-range relation modeling, an initial DSPF estimation $\boldsymbol{D}_L \in \mathbb{R}^{W_L \times H_L \times 2}$ can be derived. Starting with $\boldsymbol{D}_L$, the most straightforward strategy is to incrementally update $\boldsymbol{D}_L$ across the pyramid levels to finally achieve a high-resolution DSPF estimation $\boldsymbol{D}_1 \in \mathbb{R}^{W_1 \times H_1 \times 2}$:

$$\boldsymbol{D}_l = \mathcal{F}^{\text{DSPF}}(\boldsymbol{X}_l, \boldsymbol{X}_{l+1}^{\uparrow}, 2\boldsymbol{D}_{l+1}^{\uparrow}) \in \mathbb{R}^{W_l \times H_l \times 2}, \tag{2}$$

where $\boldsymbol{D}_{l+1}$ and $\boldsymbol{X}_{l+1}$ from a preceding level need to be upsampled in spatial resolution (denoted by "$\uparrow$"), and $\boldsymbol{D}_{l+1}$ needs to be multiplied by 2 to match the resolution of the pyramidal features in the l-th level. In this way, the estimation for DSPF can be gradually improved by fusing higher-level prediction $\boldsymbol{D}_{l+1}$ with lower-layer feature $\boldsymbol{X}_l$. Since the flow magnitude is enlarged after each estimation step, the value ranges of output spaces are different, making the regression difficult. Inspired by recent optical flow estimation techniques [20, 21], we employ a *residual refinement* strategy for more accurate DSPF estimation.

Coarse-to-Fine Residual Refinement. Rather than inferring a complete DSPF $\boldsymbol{D}_l$ at each pyramid level, our network learns the residue $\triangle\boldsymbol{D}_l \in \mathbb{R}^{W_l \times H_l \times 2}$ in relative to $\boldsymbol{D}_{l+1}$ estimated in the previous layer. Eq. (2) is thus improved as:

$$\boldsymbol{D}_l = \underbrace{\mathcal{F}^{\text{DSPF}}(\boldsymbol{X}_l, \boldsymbol{X}_{l+1}^{\uparrow})}_{\triangle\boldsymbol{D}_l} + 2\boldsymbol{D}_{l+1}^{\uparrow}. \tag{3}$$

This network design is more favored, since the magnitudes of the residues at different pyramid levels are smaller than the complete DSPFs and within a similar value range.

Recall that we apply "$\uparrow$" operation to warp a low-resolution feature map into a high-resolution one. A common way to implement such a warping operation is through bilinear upsampling. However, as bilinear upsampling only accounts for limited information from a fixed and predefined subpixel neighborhood for feature interpolation, it fails to handle the misalignment between feature maps caused by repeated padding, downsampling and upsampling, thus leading to error accumulation. Motivated by learnable feature warping operators [44, 45], we propose to learn a feature flow between two features from adjacent pyramid levels for cross-layer feature alignment.

Learnable Feature Warping. In particular, a feature flow, *i.e.*, $\boldsymbol{F}_l$, is estimated between $\boldsymbol{X}_l$ and $\boldsymbol{X}_{l+1}$:

$$\text{feature flow:} \quad \boldsymbol{F}_l = \mathcal{F}^{\text{Flow}}(\boldsymbol{X}_l, \boldsymbol{X}_{l+1}^{\Uparrow}) \in \mathbb{R}^{W_l \times H_l \times 2}. \tag{4}$$

Here, the low-resolution feature $\boldsymbol{X}_{l+1}$ is first bilinearly upsampled (denoted by "$\Uparrow$"), and then concatenated with the high-resolution feature $\boldsymbol{X}_l$ for predicting $\boldsymbol{F}_l$. The feature flow, in essence, is an offset field that aligns the positions between high-level and low-level feature maps. Through $\boldsymbol{F}_l$, $\boldsymbol{X}_{l+1}$ is warped towards $\boldsymbol{X}_l$, *i.e.*, $\boldsymbol{X}_{l+1}^{\uparrow}(\boldsymbol{u}+\boldsymbol{F}_l(\boldsymbol{u})) \sim \boldsymbol{X}_l$. Hence, we have:

$$\begin{aligned} &\text{feature warping:} \quad \boldsymbol{X}_{l+1}^{\uparrow} = \mathcal{F}^{\text{Warp}}(\boldsymbol{X}_{l+1}, \boldsymbol{F}_l) \in \mathbb{R}^{W_l \times H_l \times C_l}, \\ &\text{DSPF warping:} \quad \boldsymbol{D}_{l+1}^{\uparrow} = \mathcal{F}^{\text{Warp}}(\boldsymbol{D}_{l+1}, \boldsymbol{F}_l) \in \mathbb{R}^{W_l \times H_l \times 2}. \end{aligned} \tag{5}$$

The warping operation $\mathcal{F}^{\text{Warp}}$ upsamples the input ($\boldsymbol{X}_{l+1}$ or $\boldsymbol{D}_{l+1}$) at the $(l+1)$-th level to match the high resolution of the l-th level, according to the estimated feature flow $\boldsymbol{F}_l$. Taking the feature warping in Eq. 5 as an example, we have:

$$\boldsymbol{X}_{l+1}^{\uparrow}(\boldsymbol{u}) = \sum_{\boldsymbol{u}'_s \in \mathcal{N}(\boldsymbol{u}_s)} \boldsymbol{X}_{l+1}^{\Uparrow}(\boldsymbol{u}'_s)(1-|u_s-u'_s|)(1-|v_s-v'_s|), \tag{6}$$

where $\boldsymbol{u}_s = \boldsymbol{u} + \boldsymbol{F}_l(\boldsymbol{u}) = (u_s, v_s)$ denotes the source coordinate in $\boldsymbol{X}_{l+1}^{\Uparrow}$, $\boldsymbol{u}$ indicates the target coordinate in the interpolated output (*i.e.*, $\boldsymbol{X}_{l+1}^{\uparrow}$), and $\mathcal{N}(\boldsymbol{u}_s)$ is the four pixel neighbors of $\boldsymbol{u}_s$. Note that $\mathcal{F}^{\text{Warp}}$ is differentiable, and its gradient can be effectively computed following [23].

3.3. Fully Differentiable Human Keypoint Detection and Association

So far, we have obtained the instance-agnostic human semantic part estimation $\boldsymbol{S}$ (§3.1) and DSPF $\boldsymbol{D}$ (§3.2). Now we aim to precisely detect human joints and assemble them into human instances. By doing this, we can infer the instance-aware parsing result from $\boldsymbol{S}$, according to $\boldsymbol{D}$ (*i.e.*, the projection from dense human semantics to sparse keypoints; Fig. 3). Hence, by formulating joint association as

maximum-weight bipartite matching, a differentiable solution is derived to make our model end-to-end trainable.

Prior to pose estimation, we apply a small decoder over $\{\boldsymbol{X}_l\}_{l=1}^{L}$ to learn a pose-specific feature representation $\boldsymbol{P}$:

$$\boldsymbol{P} = \mathcal{F}^{\text{Pose}}(\{\boldsymbol{X}_l\}_{l=1}^{L}) \in \mathbb{R}^{W_1 \times H_1 \times C_1}. \quad (7)$$

$\boldsymbol{P}$ merges multi-scale cues and maintains a high resolution.

Keypoint Detection. We first predict the locations of all the visible anatomical keypoints (*e.g.*, left shoulder, right elbow) in I. Following [40, 27], we achieve this by jointly predicting confidence maps and 2D local offset fields over $\boldsymbol{P}$. Assume there are a total of K keypoint categories, and, for each category k, we compute a heatmap $M \in [0,1]^{W_1 \times H_1}$, where $M(\boldsymbol{u}) = 1$ if the location $\boldsymbol{u} = (u, v)$ falls into a disk of radius R ($= 32$ pixels) of any keypoint of category k; otherwise $M(\boldsymbol{u}) = 0$. Moreover, for each category k, we compute an offset field $\boldsymbol{O} \in \mathbb{R}^{W_1 \times H_1 \times 2}$ to improve the localization, where each offset $\boldsymbol{O}(\boldsymbol{u})$ points from $\boldsymbol{u}$ to its closest keypoint of category k. The heatmap M and offset field $\boldsymbol{O}$ are aggregated by Hough voting to obtain a highly localized Hough score map $H \in [0,1]^{W_1 \times H_1}$, whose element at location $\boldsymbol{u}$ is computed as:

$$H(\boldsymbol{u}) = \sum_{\boldsymbol{u}' \in \Omega} \frac{1}{\pi R^2} M(\boldsymbol{u}) \mathcal{F}^{\text{Bi-inter}}(\boldsymbol{u}' + \boldsymbol{O}_k(\boldsymbol{u}') - \boldsymbol{u}), \quad (8)$$

where $\mathcal{F}^{\text{Bi-inter}}$ is a bilinear interpolation function. The local maxima in H correspond to the keypoints of the k-th category. We compute K heatmaps and offset fields and conduct Hough voting individually for each category (see Fig. 4(a)). Through DSPF, we can get a set of keypoint coordinates, *i.e.*, $\{\boldsymbol{u} + \boldsymbol{D}(\boldsymbol{u})\}_{\boldsymbol{u} \in \Omega}$. With localized Hough score map of each category, we can determine the score at each of these coordinates. Finally, the coordinates whose scores are larger than 0.7 are preserved as the detected keypoints.

Limb Scoring. We denote $\mathcal{P} = \{\boldsymbol{p}_n^k : k \in \{1, \ldots, K\}, n \in \{1, \ldots, N_k\}\}$ as the set of keypoints detected in I, where N_k is the number of keypoints belonging to the k-th category, and $\boldsymbol{p}_n^k = (u_n^k, v_n^k)$ represents the 2D coordinate of the n-th detected keypoint of the k-th category. The detected joints $\mathcal{P}$ serve as candidate positions for human poses, and provide us possible kinematic connections (*i.e.*, limbs) so that we can assemble $\mathcal{P}$ to form full body poses. To do so, we first need to score each possible limb hypothesis. For each limb category, we predict a limb heatmap $Q \in \mathbb{R}^{W_1 \times H_1}$ over $\boldsymbol{P}$, which represents the confidence of the limb (see Fig. 4(a)). For simplicity, we approximately represent each limb using an elliptical area between the endpoints [29], and generate the ground-truths by applying an unnormalized elliptical Gaussian distribution with a standard deviation σ ($= 9$ pixels) over all limbs. The score of each limb is then estimated by sampling a set of Gaussian responses within the limb area, from the corresponding heatmap Q. Thus, for each limb with two joint categories k and k', we can get a scoring matrix $A \in [0,1]^{N_k \times N_{k'}}$, whose element $a_{nn'}$ stores the connectivity between joints $\boldsymbol{p}_n^k$ and $\boldsymbol{p}_{n'}^{k'}$.

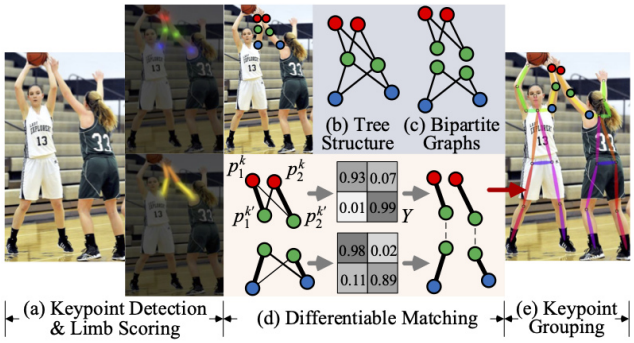


Figure 4: **Illustration of our differentiable solver for human keypoint detection and association.** See §3.3 for details.

Differentiable Matching. The keypoint candidates $\mathcal{P}$ define a large set of possible limbs. We score each limb to determine the likelihood of it belonging to a particular person, according to which we parse optimal human poses (*i.e.*, best joint matching). As shown in Fig. 4 (b), the problem of finding the optimal matching corresponds to a K-dimensional matching problem that is known to be NP-hard [51]. Thus, we decompose the matching problem into a set of bipartite matching subproblems [4], *i.e.*, determine the matching for each connected limb independently (see Fig. 4 (c)). Specifically, in each bipartite graph, a matching corresponds to a subset of edges in which no two edges share a node. Formally, we denote $\mathcal{P}^k = \{\boldsymbol{p}_n^k\}_{n=1}^{N_k}$ and $\mathcal{P}^{k'} = \{\boldsymbol{p}_{n'}^{k'}\}_{n'=1}^{N_{k'}}$ as two sets of detected keypoints belonging to category k and k', respectively. We further define a variable $y_{nn'} \in \{0,1\}$ to represent whether $\boldsymbol{p}_n^k$ and $\boldsymbol{p}_{n'}^{k'}$ are connected as a limb, and our goal becomes finding the optimal assignment $\{y_{nn'}\}_{n,n'}$ for the set of all possible connections. For each limb with joint endpoint categories k and k' and the corresponding scoring matrix $A \in [0,1]^{N_k \times N_{k'}}$, we infer the boolean assignment matrix $Y \in \{0,1\}^{N_k \times N_{k'}}$ by solving the maximum-weight bipartite matching problem:

$$\max_{Y} \sum_{n=1}^{N_k} \sum_{n'=1}^{N_{k'}} a_{nn'} \cdot y_{nn'}, \quad (9)$$

$$s.t. \quad \forall n, \quad \sum_{n'=1}^{N_{k'}} y_{nn'} \leq 1, \quad (10)$$

$$\forall n', \quad \sum_{n=1}^{N_k} y_{nn'} \leq 1, \quad (11)$$

$$\forall (n, n'), \quad y_{nn'} \in \{0,1\}, \quad (12)$$

where $a_{nn'} \in \boldsymbol{A}$ is the connection score between joints $\boldsymbol{p}_n^k$ and $\boldsymbol{p}_{n'}^{k'}$. The constraints in Eqs. (10-11) ensure that no two edges will share a node. The final multi-person pose parsing result is with maximum weight for the chosen edges (*i.e.*, connection scores). This integer linear programming problem can be solved using Hungarian algorithm [28] with a high time complexity of $O(n^3)$. Existing pose estimators [4, 39, 27] instead employ heuristic greedy algorithms, but break the end-to-end pipeline. Inspired by [61], we propose a differentiable solution which facilitates model learning with direct matching based supervision.

In particular, we drop the integer constraint (Eq.(12)) and compute Y as a real-valued assignment matrix, by solving a linear programming relaxation and rewriting Eqs. (9-12) in matrix form:

$$\min_{Y} \text{Tr}(-AY^\top), \tag{13}$$

$$s.t. \quad Y\mathbf{1}_{N_{k'}} \leq \mathbf{1}_{N_k}, \quad Y^\top \mathbf{1}_{N_k} \leq \mathbf{1}_{N_{k'}}, \quad Y \geq 0. \tag{14}$$

Here, $\mathbf{1}_{N_k} = [1]^{N_k \times 1}$ and $\mathbf{1}_{N_{k'}} = [1]^{N_{k'} \times 1}$. Since both the target function (Eq.(13)) and constraint functions (Eq.(14)) are convex, we can solve this convex constrained minimization problem using projected gradient descent (PGD) [3]. Let us denote the constraints in Eq.(14) as $\mathcal{C} = \mathcal{C}_1 \cap \mathcal{C}_2 \cap \mathcal{C}_3$, where $\mathcal{C}_1 = Y\mathbf{1}_{N_{k'}} \leq \mathbf{1}_{N_k}$, $\mathcal{C}_2 = Y^\top \mathbf{1}_{N_k} \leq \mathbf{1}_{N_{k'}}$, and $\mathcal{C}_3 = Y \geq 0$. PGD estimates Y by iterating the following equation:

$$Y \leftarrow \mathcal{F}^{\mathcal{C}}(Y - \alpha \triangledown f(Y)), \tag{15}$$

where $f(Y) = \text{Tr}(-AY^\top)$, $\triangledown f(Y) = -A$, and α indicates the step size. Here $\mathcal{F}^{\mathcal{C}}$ is a projection operator, and itself is also an optimization problem:

$$\mathcal{F}^{\mathcal{C}}(Y) = \underset{Y' \in \mathcal{C}}{\text{argmin}} \frac{1}{2}||Y' - Y||_2^2. \tag{16}$$

Given Y, $\mathcal{F}^{\mathcal{C}}$ tries to find a "point" $Y' \in \mathcal{C}$ which is closet to Y. The challenge here is how to find a feasible projection operator over the combination of several convex constraints, *i.e.*, $\mathcal{C}_1 \cap \mathcal{C}_2 \cap \mathcal{C}_3$. A popular method for finding the projection onto the intersection of convex subsets in Hilbert space is the Dykstra's cyclic projection algorithm [10, 2]. It reduces the problem to an iterative scheme which involves only finding projections from the individual sets. As the projection operator with respect to each constraint can be easily derived in a differentiable form (see *supplementary* for details), Eqs. (13-14) can be efficiently solved by PGD. In this way, our whole human part parsing model is end-to-end trainable and can direct access to keypoint matching supervision signals. After obtaining the assignment matrices for all the limbs, we follow conventions [4, 27] to sequentially group the matched keypoints from "head" to "foot", as the estimations of "head" are usually more accurate than other keypoints. Finally, multi-person pose parsing results can be delivered (see Fig. 4(d-e)).

3.4. Implementation Details

Loss Function. For category-level human semantic prediction $\boldsymbol{S}$ (§3.1), *cross-entropy* loss is used. For DSPF estimation (§3.2), l_1 loss is used. For keypoint detection and joint affinity estimation (§3.3), *cross-entropy* loss is used for confidence/score map and assignment matrix prediction, and l_1 loss is used for vector regression. The weights of different losses are adaptively adjusted during training by considering the homoscedastic uncertainty of each task [26].

Network Configuration. We compute the feature pyramid $\{\boldsymbol{X}_l\}_{l=1}^{L}$ at $L = 4$ levels with a scaling step of 2 [33]. We adopt ResNet-101 [18] as the backbone. The segmentation head $\mathcal{F}^{\text{Seg}}$ in Eq. (1) has a similar architecture as the decoder in [5], while in each upsampling stage we apply a single 5×5 depthwise-separable convolution to build a lightweight module. The pose head $\mathcal{F}^{\text{Pose}}$ in Eq. (7) is built in a similar way. For $\mathcal{F}^{\text{Flow}}$ in Eq.(4), it concatenates the inputs and then applies a 3×3 convolutional layer for feature flow prediction.

Training. We train our model in two steps. In the first step, the model is trained without differentiable matching using the SGD optimizer with a mini-batch size of 32. We start the training with a base learning rate of 5e-2, momentum of 0.9 and weight decay of 1e-5. The learning rate is then reduced to 5e-3 after 120 epochs, and the training is terminated within 150 epochs. In the second step, we finetune the whole model for another 50 epochs with a fixed learning rate of 1e-4. During all training stages, we fix the running statistics of batch normalization layers to the ImageNet-pretrained values. Typical data augmentation techniques, *e.g.*, horizontal flipping, scaling in range of [0.5, 2], rotation from [-10°, 10°], and color jitter, are also used.

Inference. Once trained, our model can be directly applied to unseen images. For a test image, we detect a set of keypoints $\mathcal{P} = \{\boldsymbol{p}_n^k\}$ and group them into different human instances. For each human instance h, we denote the set of its keypoints as $\mathcal{P}_h = \{\boldsymbol{p}_{n,h}\}_n$. Then we associate each person pixel (identified by the category-level semantic parsing module) to one of the human instances. Formally, for each human pixel $\boldsymbol{u}$ and its DSPF vector $\boldsymbol{D}(\boldsymbol{u})$, an assignment score for each human instance h is computed as:

$$\psi_h = \min_{\boldsymbol{p}_{n,h} \in \mathcal{P}_h} \left(\frac{\|\boldsymbol{u} + \boldsymbol{D}(\boldsymbol{u}) - \boldsymbol{p}_{n,h}\|_2}{(s_{n,h} + s_h) * \sigma_h}\right). \tag{17}$$

For each keypoint $\boldsymbol{p}_{n,h}$ belonging to the human instance h, its distance to $\boldsymbol{u} + \boldsymbol{D}(\boldsymbol{u})$, *i.e.*, $\boldsymbol{u}$'s associated human instance keypoint, is first computed and normalized by joint score (*i.e.*, $s_{n,h} = H(\boldsymbol{p}_{n,h})$), instance score (*i.e.*, $s_h = \frac{1}{|\mathcal{P}_h|} \sum_{\boldsymbol{p}_{n,h} \in \mathcal{P}_h} H(\boldsymbol{p}_{n,h})$), and instance scale (*i.e.*, σ_h) for reliable, scale-invariant evaluation. σ_h is set as the square root of the area of the bounding box tightly containing $\mathcal{P}_h$. Finally, the human instance for $\boldsymbol{u}$ is assigned as: $\arg\min_h \psi_h$.

4. Experiment

4.1. Experimental Setup

Datasets: Our experiments are conducted on three datasets:

- **MHP**$_{v2}$ [65] is currently the largest dataset for instance-aware human parsing. In total, it includes 25,403 images (15,403/5,000/5,000 for `train`/`val`/`test` splits). Each person is elaborately annotated with 58 fine-grained semantic categories (*i.e.*, 11 body parts, 47 cloth and accessory labels), as well as 16 body joints.
- **DensePose-COCO** [1] has 27,659 images (26,151/1,508 for `train`/`test` splits) gathered from COCO [34]. It annotates 14 human parts and 17 keypoints.
- **PASCAL-Person-Part** [53] has 1,716 and 1,817 images for `train` and `test`, respectively, with annotations of 6 human part categories and 14 body joints.

Methods	mIoU	AP^p_{50}	AP^p_{vol}	PCP_{50}	M-Adds(G)
top-down models:					
M-RCNN[17]	-	14.9	33.9	25.1	141.3
P-RCNN[58]	40.3	30.2	41.8	44.2	220.7
M-CE2P[42]	41.1	30.9	41.3	40.6	497.1
SNT[24]	-	34.4	42.5	43.5	520.5
RP-RCNN[57]	37.3	40.5	45.2	39.2	185.5
bottom-up models:					
PGN[14]	25.3	17.6	35.5	26.9	169.2
MHParser[30]	-	18.0	36.1	27.0	-
NAN[65]	-	25.1	41.8	32.3	302.2
Ours	41.4	39.0	44.3	42.3	151.1

Table 1: **Quantitative performance comparison on MHP_{v2}** `val` [65], with mIOU, AP^p and PCP. See §4.2 for details.

Methods	mIoU	AP^p_{50}	AP^p_{vol}	PCP_{50}
top-down models:				
P-RCNN[58]	65.9	43.5	53.1	51.8
M-CE2P[42]	67.1	43.7	52.9	51.2
RP-RCNN[57]	65.3	48.5	54.5	51.1
bottom-up models:				
PGN[14]	46.1	23.4	35.9	32.5
NAN[65]	58.9	37.6	48.3	43.9
Ours	69.1	49.7	54.7	52.8

Table 2: **Quantitative performance comparison on DensePose-COCO** `test` [1], with mIOU, AP^p and PCP. See §4.2 for details.

Evaluation Metrics. For instance-agnostic parsing, we adopt mean intersection-over-union (mIoU). For instance-level parsing, we employ official metrics of each dataset for fair comparison. Specifically, for MHP_{v2} and DensePose-COCO, the average precision based on parts (AP^p) and percentage of correctly parsed semantic parts (PCP) are used. For PASCAL-Person-Part, we follow conventions [16, 52] to report the average precision based on regions (AP^r).

Reproducibility. Our model is implemented in PyTorch and trained on eight NVIDIA Tesla V100 GPUs with a 32GB memory per-card. Testing is conducted on a single NVIDIA Xp GPU with 11 GB memory.

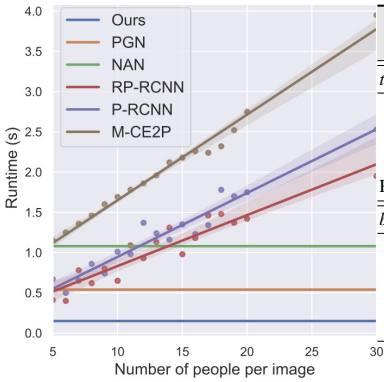
4.2. Quantitative Results

MHP_{v2} [65]: Table 1 reports comparison results against five top-down and three bottom-up approaches on MHP_{v2} `val`. Our approach significantly outperforms existing bottom-up models, *i.e.*, MHParser [30] and NAN [65], across all metrics. This indicates the importance of our multi-granularity representation learning in instance-level parsing. Furthermore, our approach shows better overall performance than most top-down methods (*e.g.*, SNT [24], P-RCNN [58], M-CE2P [42]), revealing our appealing performance. Compared to RP-RCNN [57], our approach only performs slightly worse in terms of instance-level metrics (*e.g.*, AP^p_{50}). However, our model improves the mIoU by 4% and is much more efficient (see §4.3).

DensePose-COCO [1]: Table 2 presents comparisons with five representative approaches on DensePose-COCO `val`.

Methods	AP^r_{50}	AP^r_{60}	AP^r_{70}	AP^r_{vol}
top-down models:				
MNC[8]	38.8	28.1	19.3	36.7
Li *et al.*[31]	40.6	30.4	19.1	38.4
HAZN[52]	43.7	-	-	-
P-RCNN[58]	52.9	43.0	31.1	48.5
M-CE2P[42]	53.3	45.6	31.6	51.9
RP-RCNN[57]	59.9	51.3	37.8	55.8
bottom-up models:				
PGN[14]	39.6	29.9	20.0	39.2
MH-Parser[30]	42.3	34.2	20.1	40.0
NAN[65]	59.7	51.4	38.0	52.2
Ours	59.0	52.3	38.1	55.9

Table 3: **Quantitative performance comparison on PASCAL-Person-Part** `test` [53], with AP^r. See §4.2 for details.



Methods	Speed(s) *avg. for 6 persons*
top-down models:	
SNT[24]	3.27
M-CE2P[42]	1.15
P-RCNN[58]	0.51
RP-RCNN[57]	0.43
bottom-up models:	
MHParser[30]	14.9
NAN[65]	1.08
PGN[14]	0.54
Ours	0.15

Figure 5: **Runtime analysis**. Our model is fully convolutional and allows for efficient inference, irrespective of the number of people in the image. In contrast, the runtimes of top-down approaches (*i.e.*, M-CE2P [42], P-RCNN [58], RP-RCNN [57]) grow linearly with the number of people. The runtime comparison on MHP_{v2} `val` reported in the table demonstrates again that our model is significantly faster than other methods. See §4.3 for details.

Our method sets new state-of-the-arts across all metrics, outperforming all other methods by a large margin. For example, our parser provides a considerable performance gain in AP^p_{50}, *i.e.*, 1.2% and 12.1% higher than the current best top-down (*i.e.*, RP-RCNN [57]) and bottom-up (*i.e.*, NAN [65]) models, respectively.

PASCAL-Person-Part [53]: Table 3 compares our model against six top-down and three bottom-up methods on PASCAL-Person-Part `val`. The results demonstrate that our human parser outperforms other competitors across most metrics (*i.e.*, AP^r_{60}, AP^r_{70}, AP^r_{vol}). We also observe that performance on this dataset has become saturated, due to its small-scale training set (*i.e.*, only 1,716 images).

4.3. Runtime Analysis

We collect images with a varying number of people for runtime analysis. Each analysis was repeated 500 times and then averaged. We compare with six state-of-the-art methods, including three top-down (*i.e.*, P-RCNN [58],

Figure 6: **Instance-level human semantic parsing results** (§4.4) in challenging scenarios with crowding, occlusions, pose variations, *etc*.

coarse-to-fine (Eq.(2))	residual refine. (Eq.(3))	learn. warping (Eqs.(4-6))	AP^p_{50}	AP^p_{vol}	PCP_{50}
			35.5	40.8	39.3
✓			36.9	42.4	40.3
✓	✓		38.1	43.4	41.5
✓	✓	✓	38.9	44.3	42.2

Table 4: **Ablation study on DSPF estimation** (§3.2) on MHP_{v2} `val` [65]. See §4.5 for details.

RP-RCNN [57], M-CE2P [42]) and three bottom-up (*i.e*., NAN [65], MHParser [30], PGN [14]) models. As depicted in Fig. 5, the inference times of the top-down models are roughly proportional to the number of people in the image, while they are invariant for the bottom-up approaches. The table summarizes the inference time of several parsers on MHP_{v2} `val` with average six people per image. Overall, our model is much faster than existing methods.

4.4. Qualitative Results

As shown in Fig. 6, our approach can produce precise instance-level human semantic parsing results. It shows strong robustness to many real-world challenges, such as crowding, occlusions, pose and appearance variations, *etc*.

4.5. Diagnostic Experiments

DSPF Breakdown. Table 4 investigates the necessity of our essential modules for DSPF estimation (§3.2). We start the analysis with a baseline model (*i.e*., 1_{st} row) which directly regresses DSPF over multi-layer features $\{\boldsymbol{X}_l\}_{l=1}^{L}$. Then, we progressively improve the baseline with: (1) a 'coarse-to-fine' strategy (*i.e*., inferring a complete DSPF at each pyramid level); (2)'residual refinement' (*i.e*., using residual learning with bilinear feature unsampling); and (3) 'learnable feature warping' (*i.e*., learning cross-layer feature alignment). Consistent performance improvements can be observed after introducing the above modifications and we can draw three essential conclusions: **(1)** Coarse-to-fine inference is critical for accurate DSPF regression, since it makes full use of multi-scale information. **(2)** Residual learning is also useful, since the small magnitude of the residue facilitates training. **(3)** With the learnable feature warping, cross-layer features/DSPF predictions are better aligned, further facilitating network learning.

Diff. Matching (§3.3)	AP^p_{50}	AP^p_{vol}	PCP_{50}	mIoU	mAP_{pose}
greedy matching [39]	37.6	43.1	40.8	40.8	63.9
greedy matching [4]	37.9	43.2	40.6	41.0	64.3
differentiable matching	39.0	44.3	42.3	41.4	65.8

Table 5: **Ablation study on differentiable matching** (§3.3) on MHP_{v2} `val` [65]. See §4.5 for details.

Differentiable Matching. We further study the proposed differentiable solver (§3.3) for human joint association by comparing it with two greedy matching algorithms [39, 4] on MHP_{v2} `val` (see Table 5). The differentiable solver allows directly back-propagating joint grouping errors, leading to great improvements in multi-human pose estimation (under mAP_{pose} [34]). In addition, benefiting from end-to-end learning, our model also shows better human part parsing results (in terms of AP^p_{50} and mIoU), further confirming the superiority of our differentiable matching strategy.

5. Conclusion

This work presents a new perspective of exploring human structural information over multiple granularities to address instance-aware human semantic parsing. A dense-to-sparse projection field is learnt to associate human semantic parts with human keypoints. The field is progressively optimized with residual refinement to ease the optimization. A differentiable joint matching solver is further proposed for body joint assembling. These designs together yield an end-to-end trainable, bottom-up instance-aware human semantic parser. With its accurate and fast computation, our parser is expected to pave the way for practical use and benefit vast quantities of human-centric applications.

Acknowledgment This work was in part supported by NSFC Grant 61876177, ARC DP200100938, Zhejiang Lab's Open Fund (No. 2019KD0AB04), and CCF-Baidu Open Fund.

References

[1] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 2, 6, 7

[2] James P Boyle and Richard L Dykstra. A method for finding projections onto the intersection of convex sets in hilbert spaces. In *Advances in order restricted statistical inference*, pages 28–47, 1986. 2, 6

[3] Paul H Calamai and Jorge J Moré. Projected gradient methods for linearly constrained problems. *Mathematical programming*, 39(1):93–116, 1987. 6

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 2, 3, 5, 6, 8

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 3, 6

[6] Xinlei Chen, Ross Girshick, Kaiming He, and Piotr Dollár. Tensormask: A foundation for dense object segmentation. In *ICCV*, 2019. 3

[7] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, 2018. 3

[8] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, 2016. 7

[9] Jian Dong, Qiang Chen, Xiaohui Shen, Jianchao Yang, and Shuicheng Yan. Towards unified human parsing and pose estimation. In *CVPR*, 2014. 3

[10] Richard L Dykstra. An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, 78(384):837–842, 1983. 6

[11] Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, and Song-Chun Zhu. Understanding human gaze communication by spatio-temporal graph reasoning. In *ICCV*. 2

[12] Hao-Shu Fang, Guansong Lu, Xiaolin Fang, Jianwen Xie, Yu-Wing Tai, and Cewu Lu. Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. In *CVPR*, 2018. 2

[13] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2017. 3

[14] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *ECCV*, 2018. 2, 3, 7, 8

[15] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, 2017. 1, 2

[16] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *ECCV*, 2014. 7

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 7

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[19] Shaoli Huang, Mingming Gong, and Dacheng Tao. A coarse-fine network for keypoint localization. In *ICCV*, 2017. 3

[20] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *CVPR*, 2018. 4

[21] Junhwa Hur and Stefan Roth. Iterative residual refinement for joint optical flow and occlusion estimation. In *CVPR*, 2019. 4

[22] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, 2016. 3

[23] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NeurIPS*, 2015. 4

[24] Ruyi Ji, Dawei Du, Libo Zhang, Longyin Wen, Yanjun Wu, Chen Zhao, Feiyue Huang, and Siwei Lyu. Learning semantic neural tree for human parsing. In *ECCV*, 2020. 2, 7

[25] Sheng Jin, Wentao Liu, Enze Xie, Wenhai Wang, Chen Qian, Wanli Ouyang, and Ping Luo. Differentiable hierarchical graph grouping for multi-person pose estimation. In *ECCV*, 2020. 3

[26] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018. 6

[27] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *CVPR*, 2019. 2, 3, 5, 6

[28] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 5

[29] Jia Li, Wen Su, and Zengfu Wang. Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation. In *AAAI*, 2020. 5

[30] Jianshu Li, Jian Zhao, Yunchao Wei, Congyan Lang, Yidong Li, Terence Sim, Shuicheng Yan, and Jiashi Feng. Multiple-human parsing in the wild. *arXiv preprint arXiv:1705.07206*, 2017. 2, 3, 7, 8

[31] Qizhu Li, Anurag Arnab, and Philip HS Torr. Holistic, instance-level human parsing. *arXiv preprint arXiv:1709.03612*, 2017. 1, 2, 7

[32] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan. Human parsing with contextualized convolutional neural network. In *ICCV*, 2015. 2

[33] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 6

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 6, 8

[35] Si Liu, Yao Sun, Defa Zhu, Guanghui Ren, Yu Chen, Jiashi Feng, and Jizhong Han. Cross-domain human parsing via adversarial feature and label adaptation. In *AAAI*, 2018. 2

[36] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *NeurIPS*, 2017. 3

[37] Xuecheng Nie, Jiashi Feng, and Shuicheng Yan. Mutual learning to adapt for joint human parsing and pose estimation. In *ECCV*, 2018. 1, 3

[38] Xuecheng Nie, Jiashi Feng, Jianfeng Zhang, and Shuicheng

Yan. Single-stage multi-person pose machines. In *ICCV*, 2019. 3

[39] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *ECCV*, 2018. 3, 5, 8

[40] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *CVPR*, 2017. 3, 5

[41] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*. 2

[42] 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 *AAAI*, 2019. 2, 7, 8

[43] Ziyi Shen, Wenguan Wang, Xiankai Lu, Jianbing Shen, Haibin Ling, Tingfa Xu, and Ling Shao. Human-aware motion deblurring. In *ICCV*. 2

[44] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018. 4

[45] Jiaqi Wang, Kai Chen, Rui Xu, Ziwei Liu, Chen Change Loy, and Dahua Lin. Carafe: Content-aware reassembly of features. In *ICCV*, 2019. 4

[46] Wenguan Wang, Yuanlu Xu, Jianbing Shen, and Song-Chun Zhu. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *CVPR*, 2018. 2

[47] Wenguan Wang, Zhijie Zhang, Siyuan Qi, Jianbing Shen, Yanwei Pang, and Ling Shao. Learning compositional neural information fusion for human parsing. In *ICCV*, 2019. 2

[48] Wenguan Wang, Tianfei Zhou, Siyuan Qi, Jianbing Shen, and Song-Chun Zhu. Hierarchical human semantic parsing with comprehensive part-relation modeling. *IEEE TPAMI*, 2021. 1

[49] Wenguan Wang, Hailong Zhu, Jifeng Dai, Yanwei Pang, Jianbing Shen, and Ling Shao. Hierarchical human parsing with typed part-relation reasoning. In *CVPR*. 2

[50] Fangyun Wei, Xiao Sun, Hongyang Li, Jingdong Wang, and Stephen Lin. Point-set anchors for object detection, instance segmentation and pose estimation. In *ECCV*, 2020. 2, 3

[51] Douglas Brent West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, NJ, 1996. 5

[52] Fangting Xia, Peng Wang, Liang-Chieh Chen, and Alan L Yuille. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In *ECCV*, 2016. 7

[53] Fangting Xia, Peng Wang, Xianjie Chen, and Alan L Yuille. Joint multi-person pose estimation and semantic part segmentation. In *CVPR*, 2017. 1, 2, 3, 6, 7

[54] Fangting Xia, Jun Zhu, Peng Wang, and Alan L Yuille. Pose-guided human parsing by an and/or graph using pose-context features. In *AAAI*, 2016. 1, 3

[55] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *CVPR*, 2020. 3

[56] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012. 1

[57] 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 *ECCV*, 2020. 2, 7, 8

[58] Lu Yang, Qing Song, Zhihui Wang, and Ming Jiang. Parsing R-CNN for instance-level human analysis. In *CVPR*, 2019. 2, 7

[59] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *ICCV*, 2019. 3

[60] Ze Yang, Yinghao Xu, Han Xue, Zheng Zhang, Raquel Urtasun, Liwei Wang, Stephen Lin, and Han Hu. Dense reppoints: Representing visual objects with dense point sets. In *ECCV*, 2020. 3

[61] Xiaohui Zeng, Renjie Liao, Li Gu, Yuwen Xiong, Sanja Fidler, and Raquel Urtasun. Dmm-net: Differentiable mask-matching network for video object segmentation. In *ICCV*, 2019. 2, 3, 5

[62] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *CVPR*, 2019. 3

[63] Xiaomei Zhang, Yingying Chen, Bingke Zhu, Jinqiao Wang, and Ming Tang. Part-aware context network for human parsing. In *CVPR*, 2020. 2

[64] Ziwei Zhang, Chi Su, Liang Zheng, and Xiaodong Xie. Correlating edge, pose with parsing. In *CVPR*, 2020. 1, 2

[65] 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 *ACMMM*, 2018. 2, 3, 6, 7, 8

[66] Tianfei Zhou, Siyuan Qi, Wenguan Wang, Jianbing Shen, and Song-Chun Zhu. Cascaded parsing of human-object interaction recognition. *IEEE TPAMI*, 2021. 2