Progressive One-Shot Human Parsing

Haoyu He, [1] **Jing Zhang,** [1] **Bhavani Thuraisingham,** [2] **Dacheng Tao** [1]
[1] The University of Sydney, Australia,
[2] The University of Texas at Dallas, USA
hahe7688@uni.sydney.edu.au, bxt043000@utdallas.edu, {jing.zhang1,dacheng.tao}@sydney.edu.au

Abstract

Prior human parsing models are limited to parsing humans into classes pre-defined in the training data, which is not flexible to generalize to unseen classes, $e.g.$, new clothing in fashion analysis. In this paper, we propose a new problem named one-shot human parsing (OSHP) that requires to parse human into an open set of reference classes defined by any single reference example. During training, only base classes defined in the training set are exposed, which can overlap with part of reference classes. In this paper, we devise a novel Progressive One-shot Parsing network (POPNet) to address two critical challenges , $i.e.$, ***testing bias*** and ***small sizes***. POPNet consists of two collaborative metric learning modules named Attention Guidance Module and Nearest Centroid Module, which can learn representative prototypes for base classes and quickly transfer the ability to unseen classes during testing, thereby reducing testing bias. Moreover, POPNet adopts a progressive human parsing framework that can incorporate the learned knowledge of parent classes at the coarse granularity to help recognize the descendant classes at the fine granularity, thereby handling the small sizes issue. Experiments on the ATR-OS benchmark tailored for OSHP demonstrate POPNet outperforms other representative one-shot segmentation models by large margins and establishes a strong baseline. Source code can be found at https://github.com/Charleshhy/One-shot-Human-Parsing.

Introduction

Human parsing is a fundamental visual understanding task, requiring segmenting human images into explicit body parts as well as some clothing classes at the pixel level. It has a broad range of applications especially in the fashion industry including fashion image generating (Han et al. 2019), virtual try-on (Dong et al. 2019), and fashion image retrieval (Wang et al. 2017). Although Convolutional Neural Network (CNN) has made significant progress by leveraging the large-scale human parsing datasets (Liang et al. 2015b, 2016; Ruan et al. 2019), the parsing results are restricted to the classes pre-defined in the training set, $e.g.$ 18 classes in ATR (Liang et al. 2015a), and 19 classes in LIP (Liang et al. 2018). However, due to the vast new clothing, fast varying styles in the fashion industry, parsing humans into fixed and pre-defined classes has limited the usage of human parsing models in various downstream applications.

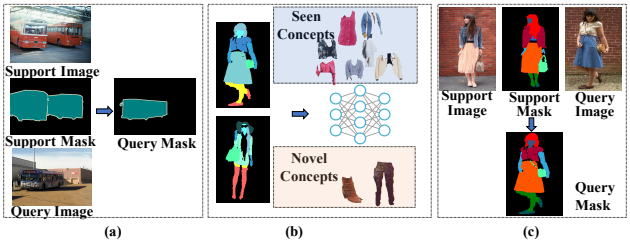

Figure 1: Comparison of OSHP against the OS3 task. (a) The classes in OS3 are large and holistic objects and only novel classes are presented and needed to be recognized during evaluation. (b) In OSHP, both base classes (cold colors) and novel (warm colors) classes are presented and needed to be recognized during the evaluation, leading to the testing bias issue. (c) In OSHP, the part of each class is small and correlated with other parts within the same human foreground.

To address the problem, we make the first attempt by defining a new task named One-Shot Human Parsing (OSHP), inspired by one-shot learning (Koch, Zemel, and Salakhutdinov 2015; Vinyals et al. 2016). OSHP requires to parse human in a query image into an open set of reference classes defined by any single reference example ($i.e.$, support image), no matter they have been seen during training (base classes) or not (novel classes). In this way, we can flexibly add and remove the novel classes depending on the requirements of specific applications without the need for collecting and annotating new training samples and retraining.

One similar task is One-shot semantic segmentation (OS3) (Zhang et al. 2019b,a; Wang et al. 2019a), that transfers the segmenting knowledge from the pre-defined base classes to the novel classes as shown in Figure 1 (a). However, OSHP is more challenging than the OS3 in two ways. Firstly, only novel classes are presented and needed to be recognized during evaluation in OS3, while both base classes and novel classes should be recognized simultaneously during evaluation in OSHP as shown in Figure 1 (b), which is indeed a variant of generalized few-shot learning (GFSL) problem (Gidaris and Komodakis 2018; Ren et al. 2019; Shi et al. 2019; Ye et al. 2019). Note that the two types

of classes have imbalanced training data, $i.e.$, there may be many training images for base classes while only a single support image for novel classes. Moreover, since we have no prior information on the explicit definition of novel classes, they are treated as background when presented during the training stage. Consequently, the parsing model may overfit the base classes and specifically lean towards the background for those novel classes, leading to the ***testing bias*** issue. Secondly, the object in OS3 to be segmented is the intact and salient foreground, while the part of each class that needs to be recognized is small and correlated with other parts within the same human foreground as shown in Figure 1(c), resulting in the ***small sizes*** issue. Directly deploying OS3 models to OSHP suffers from severe performance degradation due to these two issues.

In this work, we propose a novel POPNet for OSHP. To transfer the learning ability to recognizing base classes in the human body to the novel classes, POPNet employs a dual-metric learning strategy via an Attention Guidance Module (AGM) and Nearest Centroid Module (NCM). AGM aims to learn a discriminative feature representation for each base class ($i.e.$, prototype) while NCM is designed to enhance the transferability of such a learning ability, thereby reducing the testing bias. Although the idea of using the prototype as the class representation has been exploited in (Dong and Xing 2018; Wang et al. 2019a), we propose to gradually update them during training for the first time, which leads to learning a more robust and discriminative representation. Moreover, POPNet adopts a stage-wise progressive human parsing framework, parsing human from the coarsest granularity to the finest granularity. Specifically, it incorporates the learned parent knowledge at the coarse granularity into the learning process at the fine granularity via a Knowledge Infusion Module (KIM), which enhances the discrimination of human part features for dealing with the small sizes issue.

The main contributions of this work are as follows. Firstly, we define a new and challenging task, $i.e.$, One-Shot Human Parsing, which brings new challenges and insights to the human parsing and one-shot learning community. Secondly, to address the problem, we propose a novel one-shot human parsing method named POPNet that is composed of a dual metric learning module, a dynamic human-part prototype generator, and a hierarchical progressive parsing structure that can address the testing bias and small sizes challenges. Finally, the experiments on the ATR-OS benchmark tailored for OSHP demonstrate our POPNet achieves superior performance than representative OS3 models and can serve as a strong baseline for the new problem.

Related Work

Human Parsing

Human parsing aims at segmenting an image containing humans into semantic sub-parts including body parts and clothing classes. Recent success in deep CNN has made great progress in multiple areas (Ronneberger, Fischer, and Brox 2015; Chen et al. 2017; Zhan et al. 2020; Zhang and Tao 2020; Zhan et al. 2020; Ma et al. 2020), including human parsing (Li et al. 2017; Zhao et al. 2017; Luo et al. 2018). Instead of tackling the human parsing task with a well-defined class set, we propose to solve a new and more challenging one named OSHP, which requires to parse human into an open set of classes with only one support example. Recent methods for human parsing improve parsing performance from utilizing the body structure priors and class relations (Xiao et al. 2018; Gong et al. 2017; Zhu et al. 2018; Gong et al. 2018; Li et al. 2020; Zhan et al. 2019). One direction is modeling the parsing task together with the keypoint detection task (Xia et al. 2017; Nie et al. 2018; Huang, Gong, and Tao 2017; Fang et al. 2018; Dong et al. 2014; Zhang et al. 2020). For example, Liang $et\ al.$ proposed mutual supervision for both tasks and dynamically incorporated image-level context (Liang et al. 2018). The other direction is leveraging the hierarchical body structure at different granularities (Gong et al. 2019; He et al. 2020; Wang et al. 2019b, 2020). For example, He $et\ al.$ devised a graph pyramid mutual learning method to enhance features learned from different datasets with heterogeneous annotations (He et al. 2020). In this spirit, we also use a hierarchical structure in our POPNet to leverage the learned knowledge at the coarse granularity to aid the learning process at the fine granularity, thereby enhancing the feature representation and discrimination especially in the one-shot setting.

One-Shot Semantic Segmentation

One-Shot Semantic Segmentation (OS3) (Shaban et al. 2017) aims to segment the novel object from the query image by referring to a single support image and the support object mask. Following the one/few-shot learning (Koch, Zemel, and Salakhutdinov 2015; Finn, Abbeel, and Levine 2017; Snell, Swersky, and Zemel 2017; Sung et al. 2018; Chen et al. 2020; Liu et al. 2020a; Tian et al. 2020b), a typical OS3 solution is to learn a good metric (Zhang et al. 2018; Kate et al. 2018; Zhang et al. 2019b,a; Hu et al. 2019; Tian et al. 2020a). Zhang $et\ al.$ extracted the target class centroid and calculated the cosine similarity scores as guidance to enhance the query image features and provided a strong metric (Zhang et al. 2018). Recently, the metric was further improved by decomposing the class representations by part-aware prototypes in (Liu et al. 2020b). Besides, comparing to the one-shot one-way setting, one-shot k-way semantic segmentation has also been studied by segmenting multiple classes at the same time (Dong and Xing 2018; Wang et al. 2019a; Siam, Oreshkin, and Jagersand 2019). In contrast to the typical OS3 tasks where only intact objects of novel classes are presented and needed to be segmented, OSHP requires parsing human into small parts of both base classes and novel classes, which is similar to the challenging generalized few-shot learning (GFSL) setting tailored for practical usage scenarios (Gidaris and Komodakis 2018; Ren et al. 2019; Shi et al. 2019; Ye et al. 2019). To the best of our knowledge, GFSL for dense prediction tasks remains unexplored. In this paper, we make the first attempt by proposing a novel POPNet that employs a dual-metric learning strategy to enhance the transferability of the learning ability for recognizing human parts of base classes to novel classes.

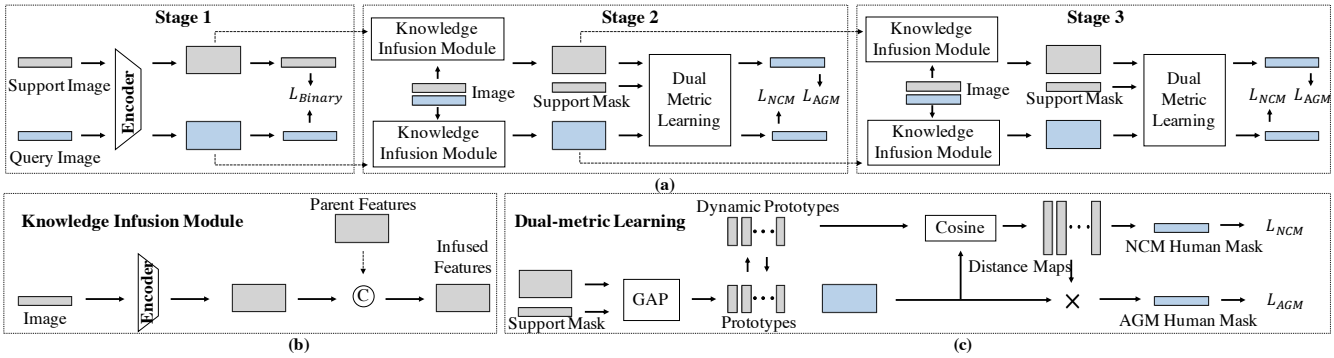

Figure 2: (a) Overview of the proposed three-stage POPNet. Each stage contains one encoder that embeds images into different semantic granularity level features. Stage 1, Stage 2, and Stage 3 generate foreground-background priors and masks, the main body area priors and masks, and the final fine-grind parsing masks respectively. (b) The structure of KIM. C denotes feature concatenation. (c) The structure of Dual-Metric Learning. GAP represents class-wise global average pooling.

Problem Definition

In this paper, we propose a new task named one-shot human parsing that requires to parse human in a query image according to the classes defined in a given support image with dense annotations. The training set is composed of many human images with dense annotations whose classes are partly overlapped with those in the support image.

Using the meta-learning language (Shaban et al. 2017; Vinyals et al. 2016; Zhang et al. 2018), in the meta-training phase, a training set $\mathcal{D}_{train}$ along with a set of classes C_{base} is given. In the meta-testing phase, the images in a test set $\mathcal{D}_{test}$ is segmented into multiple classes $C_{human} = C_{base} \cup C_{novel}$, where C_{novel} is a flexible open set of classes that have never been seen during training and can be added or removed on the fly. Specifically, both sets in $\mathcal{D}$ consist of a support set and a query set. For meta-training, the support set is denoted as $\mathcal{S}_{train} = \{(I^i, Y^i_{C_i}) | i \in [1, N_{\mathcal{S}_{train}}], C_i \subseteq C_{base}\}$, where $N_{\mathcal{S}_{train}}$ is the number of training pairs, $Y^i_{C_i}$ is the ground truth support mask annotated in $|C_i|$ human parts defined in C_i. Similarly, the query set is denoted as $\mathcal{Q}_{train} = \{(I^i, Y^i_{C_i}) | i \in [1, N_{\mathcal{Q}_{train}}], C_i \subseteq C_{base}\}$. For meta-testing, the support set $\mathcal{S}_{test}$ is similar to $\mathcal{S}_{train}$ except that the support masks are annotated according to the classes defined in C_{human}. As for the query set, only query images are provided, *i.e.*, $\mathcal{Q}_{test} = \{I^i | i \in [1, N_{\mathcal{Q}_{test}}]\}$. During the meta-training phase, training pairs (s^i, q^j) are sampled from $\mathcal{S}_{train}$ and $\mathcal{Q}_{train}$ in each episode. The meta learner aims to learn a mapping $\mathcal{F}$ subjected to $\mathcal{F}(s^i, I^j) = Y^j_{C_i}$ for any (s^i, q^j). In the meta-testing phase, the meta-learner quickly adapts the learning ability to other tasks, *i.e.*, $\mathcal{F}(s^n, I^m) = Y^m_{C_n}$ for any (s^n, q^m) sampled from $\mathcal{S}_{test}$ and $\mathcal{Q}_{test}$.

It's noteworthy that the classes in C_{test} are not necessarily connected regarding the human body structure. For example, the model can be trained with some base classes like arms and legs and evaluated with some novel classes like shoes and hat. However, we argue that given some base classes that have strong correlations with the novel ones, for example, legs and pants, it is easy to infer the novel classes like shoes in the meta-testing phase. To better evaluate the transferability of the model's learning ability to novel classes, we split the C_{train} and C_{test} to be cluster-disjoint manner which means that all the subclasses belonging to the same parent class should be in the same set. For example, C_{train} may contain hair and face (in the 'head' parent class), while C_{test} contains legs and shoes (in the 'leg' parent class). Obviously, this setting is more challenging.

Progressive One-shot Parsing Network

To address the two key challenges in OSHP, we devise a POPNet. It has three stages of different granularity levels (Figure 2 (a)). The first stage generates the foreground body masks, the second stage generates the coarse main body area masks, and the third stage generates the final fine-grind parsing masks. The learned semantic knowledge of each stage is inherited by the next stage via a Knowledge Infusion Module (Figure 2 (b)) to enhance the discrimination of human part features and deal with the small sizes issue. In the second and third stage, a Dual-metric Learning (DML)-based meta-learning method ((Figure 2 (c))) is proposed to generate robust dynamic class prototypes that can generalize to the novel classes thereby reducing the testing bias.

Progressive Human Parsing

Instead of being intact objects, human parsing classes are non-holistic small human parts, which makes it non-trivial to directly adopt the OS3 metric learning methods on the OSHP task. Inspired by (He et al. 2020), the human body is highly structural and the semantics at human coarse granularity can help network eliminate distractions and focus on the target classes at the fine granularity. To this end, we decompose our POPNet into three stages from the coarse granularity to the fine granularity. By infusing the learned knowledge from the coarse stages into the fine stage via a knowledge infusion module (detailed in Section), the network boosts the pixel-wise feature representations with rich parents semantics to discriminate the small-sized human parts.

Specifically, in one episode, we are provided with the query image $I^q \in \mathbb{R}^{H \times W \times 3}$, support image $I^s \in \mathbb{R}^{H \times W \times 3}$

and support mask $Y^s_{C_s} \in \mathbb{R}^{H\times W\times|C_s|}$ that annotated in class set C_s. The network's expected outcome is the predicted $Y^q_{C_s} \in \mathbb{R}^{H\times W\times|C_s|}$ which assigns the classes in the support mask to the query image pixels. In the first stage, we devise a binary human parser that can segment the human foreground out of the background. It is trained via supervised learning by leveraging additional binary masks $Y^q_{C_{fg}}$ and $Y^s_{C_{fg}}$, $i.e.$, $|C_{fg}| = 2$, derived from Y^q_C and Y^s_C by replacing all the foreground classes with a single foreground label. Noting that here we adopt the conventional supervised foreground segmentation setting instead of the one-shot one-way foreground segmentation setting since a well-trained human foreground parser can include most of the possible human-related classes and cause no harm to the potential novel classes semantically. Besides, there is no large-scale one-shot segmentation dataset that contains the human class while having a small domain gap with the existing human parsing datasets. We leave it as the future work to explore the one-shot one-way setting in the first stage.

In the second stage, we follow the OSHP settings and devise a one-shot meta learner on the main body areas that the parsed foreground classes in this stage are at the coarse granularity, $i.e.$, head, body, arms, legs, and the background. Assuming $C_{\hat{s}}$ is the set of the main body areas, we can get the supervision $Y^s_{C_{\hat{s}}}$ by aggregating $Y^s_{C_s}$, $i.e.$, replacing the class labels belonging to the same parent class with the parent class label. Hence, $|C_{\hat{s}}| = 4$ and $|C_{\hat{s}}| = 5$ during the meta-training and meta-testing respectively, since one coarse class serves as the novel class during training according to Section. . Accordingly, the model learns the body semantics via the meta-learning in the second stage. In the third stage, we devise the one-shot human parsing meta learner that can predict the fine-granularity human classes $Y^q_{C_s}$.

Knowledge Infusion Module

To fully exploit context information from the the previous stages and enhance the representative ability of the features in the current stage, we propose to infuse the learned parent knowledge into learning process when inferring its descendants. Specifically, in each stage, the input image (query image or support image) is fed into a shared encoder network to get the embedded features $g^{S_i}, i = 1, 2, 3$. In the second stage, we exploit g^{S_1} by concatenating it with the image features learned in the second stage g^{S_2} to get the enhanced features h^{S_2} via a knowledge infusion module, $i.e.$, $h^{S_2} = \zeta_2\left([g^{S_1}; g^{S_2}]\right)$, where $[;]$ denotes the concatenation operator, ζ_2 represents the mapping function learned by two consecutive conv layers. Likewise, in the third stage, we exploit h^{S_2} by concatenating it with the image features learned in the third stage g^{S_3} to get the enhanced features h^{S_3}, $i.e.$, $h^{S_3} = \zeta_3\left([h^{S_2}; g^{S_3}]\right)$. All encoded features and infused features are in $\mathbb{R}^{H\times W\times K}$, where K denotes feature channels.

We implement the encoder in each stage using the Deeplab v3+ model (Chen et al. 2018) with an Xception backbone (Chollet 2017). We use the features before the classification layer as g^{S_i}, since they contain semantic-related information. In this way, the learned hierarchical body structure knowledge is infused into the next stage progressively to help discriminate the fine-granularity classes via dual-metric learning (detailed in Section). We train the three stages sequentially and fix the model parameters in the previous stage when training the current stage.

Dual-Metric Learning

Take the third stage as an example, given the support mask $Y^s_{C_s}$, infused features h^s and h^q, it is desired to generate the query mask $Y^q_{C_s}$. In the OS3 methods, inferring the query mask is accomplished by using convolution layers (Hu et al. 2019; Gairola et al. 2020; Zhang et al. 2019b) or graph reasoning (Zhang et al. 2019a) to explore pixel relationships. Recently, (Tian et al. 2020b) propose to solve support-query inconsistency by enriching the features through a pyramid-alike structure. The mentioned approaches can be summarized as post feature enhancement approaches that learn the implicit query-support correlations after the encoder. However, the learned query-support correlations in the enriched features are likely to be overfitting on the base class, thereby reducing the transferability on the generalized OSHP setting. In this case, we choose the simple yet effective design that computes the cosine similarity scores (Zhang et al. 2018; Liu et al. 2020b) between the features and class prototypes, which shows a better transferability.

Dynamic Prototype Generation Different to the prior prototype methods, we propose to generate more robust dynamic prototypes. First, we calculate the class prototype p_c for class $c \in C_s \setminus c_{bg}$ (Zhang et al. 2018) as:

$$p_c = \frac{1}{|\Lambda_c|}\sum_{(x,y)\in\Lambda_c} h^s\left(x, y\right), \tag{1}$$

where (x, y) denote pixel index, Λ_c is the support mask of class c, $|\Lambda_c|$ is the number of pixels in the mask. Note that in the prior methods (Dong and Xing 2018) that a 'background prototype' is learned to represent non-foreground regions. However, in the OSHP setting, the background pixels in the training data include both background and the novel classes in C_{novel}. Therefore, we do not calculate the background prototype to prevent pushing the novel classes towards background class. Instead, we predict the background by excluding all the foreground classes in the following sessions.

Instead of using a static p_c in the following networks, we generate a dynamic prototype p^d_c to improve the robustness of base class representation. Specifically, it is calculated by gradually smoothing the previous prototype estimate p^d_c and the current estimate p_c in each episode, $i.e.$,

$$p^d_c = \alpha \times p^d_c + (1 - \alpha) \times p_c, \tag{2}$$

where α is the smoothing parameter. Since the novel classes are not seen in training, we use static prototypes for the novel classes during testing. For simplicity, we denote both prototypes as $\hat{p}$ in the following sessions. Next, the distance map m_c between the query features h_q and the class prototype is calculated by cosine similarity as follows: $m_c =< h^q, \hat{p}_c >$. Prior methods mainly utilize distance maps in two ways. Parametric approach (Zhang et al. 2018) uses distance maps as the attention by element-wise multiplying the distance

map to the query image feature maps for further prediction. Non-parametric approach (Wang et al. 2019a) directly makes predictions basing on the distance maps. We find that the first approach can learn a better metric on the base classes while cannot generalize well on the novel classes. In contrast, the second approach has a strong transferability due to the effective and simple distance metric, but it struggles to discriminate the human classes that are semantically similar. To this end, we propose a novel weight shifting strategy for DML such that it disentangles metric's representation ability and model's generalization ability. In the early training phase, DML learns the metric for better representation using AGM. In the late phase, DML shifts focus to improve the transferability of this learning ability on novel classes and addresses testing bias issue using NCM.

Attention Guidance Module In the early training phase, our meta learner aims to fully exploit the supervisory signals from base classes and learn a good feature representation. To this end, we use the distance maps m_c as the class-wise attention to enhance the query features in a residual learning way, $i.e.$, $r_c = m_c \times h^q + h^q$. Then, we generate probability map for each class by feeding r_c to a convolutional layer φ and a softmax layer, $i.e.$,

$$\begin{aligned} Y_c^{q;AGM} &= \frac{exp(\varphi(r_c))}{\sum_{c \in C_s \setminus c_{bg}} exp(\varphi(r_c)) + r_{bg}} \\ r_{bg} &= (1/(|C_s| - 1)) \times \sum_{c \in C_s \setminus c_{bg}} \omega(r_c). \end{aligned} \tag{3}$$

Note that we infer the probability map for the background class by aggregating all the foreground features after a convolutional layer ω, which can automatically attend to the non-foreground regions by learning negative weights.

Nearest Centroid Module In the late training phase, our meta learner aims to increase the transferability of the learning ability from base classes to novel classes. To this end, we propose the non-parametric Nearest Centroid Module that infers the probability map directly from the similarity between features and class prototypes. Specifically, we use a softmax layer directly on the distance maps m_c and m_{bg} to get the final prediction. Likewise, we get m_{bg} by explicitly averaging all the reverse foreground distance maps, $i.e.$,

$$\begin{aligned} Y_c^{q;NCM} &= \frac{exp(m_c)}{\sum_{c \in C_s \setminus c_{bg}} exp(m_c) + m_{bg}} \\ m_{bg} &= (1/(|C_s| - 1)) \times \sum_{c \in C_s \setminus c_{bg}} (1 - m_c). \end{aligned} \tag{4}$$

Weight Shifting Strategy During training, we control the meta-learner's focus by assigning dynamic loss weights for both modules, $i.e.$,

$$\begin{aligned} L = \beta \times (-\frac{1}{N} \sum_{x,y,z} \mathbb{I}_{c=t} log(y_c^{q;AGM}(x,y))) \\ + (1-\beta) \times (-\frac{1}{N} \sum_{x,y,z} \mathbb{I}_{c=t} log(y_c^{q;NCM}(x,y))), \end{aligned} \tag{5}$$

where $\mathbb{I}_{c=t}$ is a binary indicator function outputting 1 when class c is the target class. β denotes the loss weight that decreases with the increases of training epoch, $i.e.$, $\beta = 1 - epoch/max\_epoch$, thereby gradually shifting the meta-learner's focus from AGM to NCM.

Dataset and Metric

Dataset: ATR-OS In this session, we illustrate how to tailor the existing large-scale ATR dataset (Liang et al. 2015a,b) into a new ATR-OS dataset for the OSHP setting. We choose the ATR dataset instead of the MHP dataset (Li et al. 2017; Zhao et al. 2018) for the following reasons. First, ATR dataset a large-scale benchmark including 18000 images annotated with 17 foreground classes. The abundant labeled data allow the network to learn rich feature representations. Second, ATR's images are mostly fashion photographs including models and a variety of fashion items, which are closely related to OSHP's applications such as fashion clothing parsing (Yamaguchi et al. 2012). Third, comparing to the other datasets, models' poses, sizes, and positions in the ATR dataset have less diversity. Hence, it is a good start for the newly proposed challenging OSHP task. We leave the research on OSHP in complex scenes as future work.

We split the ATR samples into support sets and query sets according to the one-shot learning setting for training and testing respectively. We form $\mathcal{Q}_{train}$ by including the first 8000 images of the ATR training set and form $\mathcal{S}_{train}$ with the remaining images. We form the 500-image $\mathcal{Q}_{test}$ and 500-image $\mathcal{S}_{test}$ from the original test set in a similar way. In each training episode, we randomly select one query-support pair from $\mathcal{Q}_{train}$ and $\mathcal{S}_{train}$, while in each testing episode, the network is evaluated by mapping each sample from $\mathcal{Q}_{test}$ to 10 support samples from $\mathcal{S}_{test}$ and forms a 5000 testing pairs in total. For a fair comparison, the 10 support samples are fixed. The selection for $\mathcal{S}_{test}$ images is illustrated in supplementary materials.

To ease the difficulty for training OSHP on the ATR dataset, we merge the symmetric classes and rare classes in ATR, $e.g.$ 'left leg' and 'right leg' are merged as 'legs' and 'sunglasses' is merged into the background. Before training, the remaining 12 classes including 'background' denoted as C_{human} are sampled into C_{base} and C_{novel}. To limit the networks to only learn from the classes in C_{base} during training, the regions of C_{novel} are merged into 'background', thereby only classes in C_{base} are seen in D_{train}. During testing, all classes indicated by the support masks are evaluated, including classes from both C_{base} and C_{novel}. Note that it is unreasonable in a query-support pair that some classes required to be parsed in the query image are not annotated in the support mask, so we merge these classes into 'background' as well. Besides, due to the reason illustrated in Section , C_{novel} is chosen from the two sets representing two main body areas, respectively, $i.e.$, the leg area: $C_{Fold\ 1}$ = [pants, legs, shoes] and the head area: $C_{Fold\ 2}$ = [hair, head, hat].

Metrics We use Mean Intersection over Union (MIoU) as the main metric for evaluating the parsing performance on the novel classes C_{novel} and all the human classes C_{human}. We also compute average overall accuracy to evaluate the overall human parsing performance. For the one-way setting as described in Section , we also compute the average Binary-IoU (Wang et al. 2019a). To avoid confusion, we refer to the main evaluation setting as k-way OSHP that parses k human parts at the same time while we refer to parsing only one class in each episode as one-way OSHP.

Method	Novel Class MIoU			Human MIoU			Overall Acc
	Fold 1	Fold 2	Mean	Fold 1	Fold 2	Mean	
AMP	8.5	8.1	8.3	16.3	15.4	15.9	67.6
SG-One	0.0	0.1	0.1	42.7	46.0	44.4	91.6
PANet	12.6	13.3	13.0	19.4	17.1	18.3	78.8
POPNet	**24.1**	**19.4**	**21.8**	**60.6**	**60.4**	**60.5**	**94.1**

Table 1: Comparison on k-way OSHP with the baselines. Human MIoU refers to the MIoU on C_{human}.

Method	Novel Class MIoU			Human MIoU			Bi-IoU
	Fold 1	Fold 2	Mean	Fold 1	Fold 2	Mean	
Fine-tune	0.3	0.2	0.3	14.8	15.0	14.9	49.1
AMP	8.4	9.4	8.8	15.0	15.0	15.0	50.7
SG-One	4.0	0.7	2.4	39.0	40.5	39.8	66.0
PANet	5.1	3.2	4.2	14.0	13.9	14.0	49.5
POPNet	**28.3**	**28.4**	**27.7**	**51.1**	**54.6**	**52.8**	**71.4**

Table 2: Comparison on one-way OSHP with the baselines.

Experiments

Baselines

Fine-Tuning: as suggested in (Caelles et al. 2017), we first pre-train the model on $\mathcal{D}_{train}$ then fine-tune on the $\mathcal{S}_{test}$ for a few iterations. Specifically, we use the same backbone as POPNet and only fine-tune the last two convolution layers and the classification layer. **SG-One**: we follow the settings of SG-One (Zhang et al. 2018) and learn similarity guidance from the support image features and support mask. We use the same backbone as POPNet for better performance. To support k-way OSHP, we follow a similar prediction procedure as defined by Eq. (3) in our AGM except that it does not use residual learning. **PANet**: we use PANet as another baseline with non-parameter metric learning and prototype alignment loss. In the k-way OSHP, we pair each query image with k support images that each contains a unique class ($i.e.$, with a binary support mask) in the support set as is described in (Wang et al. 2019a). **AMP**: we use masked proxies with multi-resolution weight imprinting technology and carefully tune a suitable learning rate as described in (Siam, Oreshkin, and Jagersand 2019).

Implementation Details

In this paper, we conduct the experiments on a single NVIDIA Tesla V100 GPU. The backbone network is pre-trained on the COCO dataset (Lin et al. 2014). The images are resized to 576×576 in one-way OSHP tasks and resized to 512×512 in k-way tasks due to memory limit, which leads to computations of 131.3 GMacs and 100.4 GMacs respectively. Training images are augmented by a random scale from 0.5 to 2, random crop, and random flip. We train the model using the SGD optimizer for 30 epochs with the poly learning rate policy. The initial learning rate is set to 0.001 with batch size 2. When generating dynamic prototypes, α in Eq. (2) is set to 0.001 by grid search. However, static prototypes are utilized when calculating distance maps in the first 15 epochs before we aggregating enough prototypes to reduce the variance and get stable dynamic prototypes.

Figure 3: Visual results on ATR-OS. (a) One-way OSHP. (b) K-way OSHP.

Results and Analysis

Comparison With Baselines **K-Way OSHP**: we compare our model with the customized OS3 baseline models described in Table 1. We first report the results on the overall human classes, POPNet significantly improves the Human MIoU and Overall Accuracy to 60.5% and 94.1% respectively, which demonstrates the three-stage progressive parsing can develop pixel-wise feature representations with rich semantic that address the small-sized human part issue. When evaluating the novel classes, our method has outperformed the baseline methods by 8.8%. The significant margin demonstrates that the class centroid learned by dynamic prototypes in DML can successfully be generalized to novel classes and reduce the testing bias effect.

One-Way OSHP: in addition to k-way OSHP, we also report results on one-way OSHP that the support mask only contains one class in one episode. In this setting, we evaluate each class 500 times with different (s, q) pairs randomly sampled from $\mathcal{S}$ and $\mathcal{Q}$. As is seen from Table 2, our model achieves 27.7%, 52.8%, and 71.4% in mean novel class MIoU, mean human MIoU, and mean Bi-IoU and outperforms the best baseline method by a margin of 18.9%, 13.0%, and 5.4% respectively. POPNet's superiority in solving small human parts and novel classes is again confirmed by the large margins on the one-way OSHP. Note that when comparing one-way OSHP scores to k-way OSHP, the novel class MIoU is higher while the overall human MIoU is lower. The reason is that the model would be less confident in the novel classes when the base classes are involved in the testing at the same time. However, multiple classes' prototypes from the k-way supervisory signals would help our model learn the underlying human part semantic relations and improve the overall human parsing performance.

Figure 4: Visual comparison with baselines on ATR-OS.

Methods	Novel Class MIoU	Human MIoU
AGM	1.0	40.2
NCM	3.5	40.4
DML	15.1	47.9
DML + WS	18.6	47.1
DML + WS + KIM	20.0	52.5
DML + WS + KIM + DP	24.1	60.6

Table 3: Ablation study on ATR-OS Fold 1. WS is short for the weight shifting strategy, KIM is short for the three-stage progressive human parsing with knowledge infusion module, and DP is short for the dynamic prototype generation.

Visual Inspection To better understand the OSHP task and POPNet, we show the qualitative results on both OSHP settings in Figure 3. In one-way OSHP, when given image-mask support pair on one novel class, $e.g.$ 'shoes' in Fold 1, POPNet can segment the part mask of the novel class from the query image accurately although the appearance gap is huge. In k-way OSHP, when the full human mask is given, POPNet can efficiently parse human into multiple human parts including both base and novel classes. The qualitative results show that our method can flexibly generate satisfying parsing masks with classes defined by the support example.

We further compare the visual results with the baseline methods on ATR-OS Fold 1 in Figure 4. As is seen, the baseline methods struggle in recognizing the small-sized human parts and are only able to separate the holistic human foreground from the background. Although SG-One (Zhang et al. 2018) can segment the pixels of base classes, $e.g.$ 'hair', it tends to overfit these base classes and cannot find the novel classes. In contrast, POPNet can discriminate and parse the non-holistic small classes by reducing the testing bias effect via DML and tackling the small-sized human parts with the progressive three-stage structure.

Ablation Study We investigate the effectiveness of key POPNet components in this session. Firstly, as in Table 3, when applying either AGM or NCM on the ATR-OS dataset ($\beta = 0$ or $\beta = 1$ in Eq (5)), the two models only achieve less than 5% novel class MIoU and around 40% human MIoU. The scores suggest that the models can only recognize the holistic human contour and cannot segment the novel classes due to the small sizes and testing bias challenges. Next, we evaluate the weight shifting effect by comparing DML with and without weight shifting. DML without weight shifting ($\beta = 0.5$ in Eq (5)) utilizes both metric learning modules and achieves 15.1% novel class MIoU and 47.9% human MIoU, much better than any single module. By using the weight-shifting strategy, our model can significantly reduce the testing bias, improving the novel class MIoU by a margin of 3.5%. It validates that shifting the network's focus on NCM during the late stage of training can considerably improve the network's transferability to novel classes. We then apply progressive human parsing on the existing structure, the human MIoU is noticeably improved by 5.4%, and novel class MIoU is improved to 20%, which demonstrates that employing hierarchical human structure is beneficial for tackling the small-sized human parts. Finally, POPNet can achieve 24.1% novel class MIoU and 60.6% human MIoU through dynamic prototype generation. In addition to the table content, the base class MIoU from $C_{human} \setminus C_{novel}$ reaches 72.8%, which is close to the fully supervised human parsing methods. It is indicated that increasing the robustness of the base class representation is also helpful for transferring the knowledge to the novel classes and gaining a remarkable margin on the novel class MIoU.

Limitation We find that there is still a performance gap between the novel classes and the base classes. Such gap mainly comes from the confusion among the novel classes, $e.g.$ shoes, and legs in Figure 3 (a). To address this issue, modeling the inter-class relations using graph neural network and reasoning on the graph may enhance the feature discrimination further, which we leave as the future work.

Conclusion

We introduce a new challenging but promising problem, $i.e.$, one-shot human parsing, which requires parsing human into an open set of classes defined by a single reference image. Moreover, we make the first attempt to build a strong baseline, $i.e.$, Progressive One-shot Parsing Network (POPNet). POPNet adopts a dual-metric learning strategy based on dynamic prototype generation, which demonstrates its effectiveness for transferring the learning ability from base seen classes to novel unseen classes. Moreover, the progressive parsing framework effectively leverages human part knowledge learned at the coarse granularity to aid the feature learning at the fine granularity, thereby enhancing the feature representations and discrimination for small human parts. We also tailor the popular ATR dataset to the one-shot human parsing settings and compare POPNet with other representative one-shot semantic segmentation models. Experiment results confirm that POPNet outperforms other models in terms of both generalization ability on the novel classes and the overall parsing ability on the entire human body. The future work may include 1) constructing new benchmarks for this task by paying more attention to appearance diversity, $e.g.$, pose and occlusion; and 2) modeling inter-class correlations to enhance the feature representations.

Acknowledgements

This work was supported by the Australian Research Council Projects FL-170100117, DP-180103424, IH-180100002.

References

Caelles, S.; Maninis, K.-K.; Pont-Tuset, J.; Leal-Taixé, L.; Cremers, D.; and Van Gool, L. 2017. One-shot video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 221–230.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. 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.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision*, 801–818.

Chen, Y.; Wang, X.; Liu, Z.; Xu, H.; and Darrell, T. 2020. A New Meta-Baseline for Few-Shot Learning. *arXiv preprint arXiv:2003.04390* .

Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 1251–1258.

Dong, H.; Liang, X.; Shen, X.; Wu, B.; Chen, B.-C.; and Yin, J. 2019. Fw-gan: Flow-navigated warping gan for video virtual try-on. In *Proceedings of the IEEE International Conference on Computer Vision*, 1161–1170.

Dong, J.; Chen, Q.; Shen, X.; Yang, J.; and Yan, S. 2014. Towards unified human parsing and pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 843–850.

Dong, N.; and Xing, E. 2018. Few-Shot Semantic Segmentation with Prototype Learning. In *Proceedings of the British Machine Vision Conference*.

Fang, H.-S.; Lu, G.; Fang, X.; Xie, J.; Tai, Y.-W.; and Lu, C. 2018. Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. *arXiv preprint arXiv:1805.04310* .

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, 1126–1135.

Gairola, S.; Hemani, M.; Chopra, A.; and Krishnamurthy, B. 2020. SimPropNet: Improved Similarity Propagation for Few-shot Image Segmentation. In *International Joint Conference on Artificial Intelligence*.

Gidaris, S.; and Komodakis, N. 2018. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4367–4375.

Gong, K.; Gao, Y.; Liang, X.; Shen, X.; Wang, M.; and Lin, L. 2019. Graphonomy: Universal Human Parsing via Graph Transfer Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7450–7459.

Gong, K.; Liang, X.; Li, Y.; Chen, Y.; Yang, M.; and Lin, L. 2018. Instance-level human parsing via part grouping network. In *Proceedings of the European Conference on Computer Vision*, 770–785.

Gong, K.; Liang, X.; Zhang, D.; Shen, X.; and Lin, L. 2017. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 932–940.

Han, X.; Hu, X.; Huang, W.; and Scott, M. R. 2019. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE International Conference on Computer Vision*, 10471–10480.

He, H.; Zhang, J.; Zhang, Q.; and Tao, D. 2020. Grapy-ML: Graph Pyramid Mutual Learning for Cross-Dataset Human Parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 10949–10956.

Hu, T.; Yang, P.; Zhang, C.; Yu, G.; Mu, Y.; and Snoek, C. G. 2019. Attention-based multi-context guiding for few-shot semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 8441–8448.

Huang, S.; Gong, M.; and Tao, D. 2017. A coarse-fine network for keypoint localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 3028–3037.

Kate, R.; Evan, S.; Trevor, D.; Alyosha A., E.; and Sergey, L. 2018. Conditional Networks for Few-Shot Semantic Segmentation. In *ICLR workshop*.

Koch, G.; Zemel, R.; and Salakhutdinov, R. 2015. Siamese neural networks for one-shot image recognition. In *ICML workshop*.

Li, J.; Zhao, J.; Wei, Y.; Lang, C.; Li, Y.; Sim, T.; Yan, S.; and Feng, J. 2017. Multiple-human parsing in the wild. *arXiv preprint arXiv:1705.07206* .

Li, T.; Liang, Z.; Zhao, S.; Gong, J.; and Shen, J. 2020. Self-learning with rectification strategy for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9263–9272.

Liang, X.; Gong, K.; Shen, X.; and Lin, L. 2018. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(4): 871–885.

Liang, X.; Liu, S.; Shen, X.; Yang, J.; Liu, L.; Dong, J.; Lin, L.; and Yan, S. 2015a. Deep human parsing with active template regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(12): 2402–2414.

Liang, X.; Shen, X.; Feng, J.; Lin, L.; and Yan, S. 2016. Semantic object parsing with graph lstm. In *Proceedings of the European Conference on Computer Vision*, 125–143.

Liang, X.; Xu, C.; Shen, X.; Yang, J.; Liu, S.; Tang, J.; Lin, L.; and Yan, S. 2015b. Human parsing with contextualized convolutional neural network. In *Proceedings of the IEEE International Conference on Computer Vision*, 1386–1394.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, 740–755.

Liu, W.; Zhang, C.; Lin, G.; and Liu, F. 2020a. CRNet: Cross-Reference Networks for Few-Shot Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4165–4173.

Liu, Y.; Zhang, X.; Zhang, S.; and He, X. 2020b. Part-aware Prototype Network for Few-shot Semantic Segmentation. In *Proceedings of the European Conference on Computer Vision*.

Luo, Y.; Zheng, Z.; Zheng, L.; Guan, T.; Yu, J.; and Yang, Y. 2018. Macro-micro adversarial network for human parsing. In *Proceedings of the European Conference on Computer Vision*, 418–434.

Ma, B.; Zhang, J.; Xia, Y.; and Tao, D. 2020. Auto Learning Attention. In *Advances in Neural Information Processing Systems*.

Nie, X.; Feng, J.; Zuo, Y.; and Yan, S. 2018. Human pose estimation with parsing induced learner. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2100–2108.

Ren, M.; Liao, R.; Fetaya, E.; and Zemel, R. 2019. Incremental few-shot learning with attention attractor networks. In *Advances in Neural Information Processing Systems*, 5275–5285.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, 234–241.

Ruan, T.; Liu, T.; Huang, Z.; Wei, Y.; Wei, S.; and Zhao, Y. 2019. Devil in the details: Towards accurate single and multiple human parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4814–4821.

Shaban, A.; Bansal, S.; Liu, Z.; Essa, I.; and Boots, B. 2017. One-shot learning for semantic segmentation. In *Proceedings of the British Machine Vision Conference*.

Shi, X.; Salewski, L.; Schiegg, M.; Akata, Z.; and Welling, M. 2019. Relational generalized few-shot learning. *arXiv preprint arXiv:1907.09557* .

Siam, M.; Oreshkin, B. N.; and Jagersand, M. 2019. AMP: Adaptive masked proxies for few-shot segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 5249–5258.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, 4077–4087.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1199–1208.

Tian, P.; Wu, Z.; Qi, L.; Wang, L.; Shi, Y.; and Gao, Y. 2020a. Differentiable Meta-Learning Model for Few-Shot Semantic Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 12087–12094.

Tian, Z.; Zhao, H.; Shu, M.; Yang, Z.; Li, R.; and Jia, J. 2020b. Prior Guided Feature Enrichment Network for Few-Shot Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. In *Advances In Neural Information Processing Systems*, 3630–3638.

Wang, K.; Liew, J. H.; Zou, Y.; Zhou, D.; and Feng, J. 2019a. Panet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, 9197–9206.

Wang, W.; Zhang, Z.; Qi, S.; Shen, J.; Pang, Y.; and Shao, L. 2019b. Learning Compositional Neural Information Fusion for Human Parsing. In *Proceedings of the IEEE International Conference on Computer Vision*, 5703–5713.

Wang, W.; Zhu, H.; Dai, J.; Pang, Y.; Shen, J.; and Shao, L. 2020. Hierarchical human parsing with typed part-relation reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Wang, Z.; Gu, Y.; Zhang, Y.; Zhou, J.; and Gu, X. 2017. Clothing retrieval with visual attention model. In *2017 IEEE Visual Communications and Image Processing*, 1–4.

Xia, F.; Wang, P.; Chen, X.; and Yuille, A. L. 2017. Joint multi-person pose estimation and semantic part segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6769–6778.

Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; and Sun, J. 2018. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision*, 418–434.

Yamaguchi, K.; Kiapour, M. H.; Ortiz, L. E.; and Berg, T. L. 2012. Parsing clothing in fashion photographs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3570–3577.

Ye, H.-J.; Hu, H.; Zhan, D.-C.; and Sha, F. 2019. Learning Adaptive Classifiers Synthesis for Generalized Few-Shot Learning. *arXiv preprint arXiv:1906.02944* .

Zhan, Y.; Yu, J.; Yu, T.; and Tao, D. 2019. On exploring undetermined relationships for visual relationship detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5128–5137.

Zhan, Y.; Yu, J.; Yu, T.; and Tao, D. 2020. Multi-task Compositional Network for Visual Relationship Detection. *International Journal of Computer Vision* 128(8): 2146–2165.

Zhang, C.; Lin, G.; Liu, F.; Guo, J.; Wu, Q.; and Yao, R. 2019a. Pyramid Graph Networks with Connection Attentions for Region-Based One-Shot Semantic Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 9587–9595.

Zhang, C.; Lin, G.; Liu, F.; Yao, R.; and Shen, C. 2019b. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5217–5226.

Zhang, J.; and Tao, D. 2020. Empowering Things with Intelligence: A Survey of the Progress, Challenges, and Opportunities in Artificial Intelligence of Things. *IEEE Internet of Things Journal* .

Zhang, X.; Wei, Y.; Yang, Y.; and Huang, T. 2018. Sg-one: Similarity guidance network for one-shot semantic segmentation. *arXiv preprint arXiv:1810.09091* .

Zhang, Z.; Su, C.; Zheng, L.; and Xie, X. 2020. Correlating Edge, Pose with Parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zhao, J.; Li, J.; Cheng, Y.; Sim, T.; Yan, S.; and Feng, J. 2018. 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*, 792–800.

Zhao, J.; Li, J.; Nie, X.; Zhao, F.; Chen, Y.; Wang, Z.; Feng, J.; and Yan, S. 2017. Self-supervised neural aggregation networks for human parsing. In *CVPR workshop*, 7–15.

Zhu, B.; Chen, Y.; Tang, M.; and Wang, J. 2018. Progressive Cognitive Human Parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 7607–7614.