UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Data-efficient Visual Understanding via Deep Neural Networks

by Peike Li

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Dr. Xin Yu

University of Technology Sydney
Faculty of Engineering and Information Technology

July 2022

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Peike Li* declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the *Faculty of Engineering and Information Technology* at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

SIGNATURE: Production Note: Signature removed prior to publication.

DATE: 11th July, 2022

PLACE: Sydney, Australia

ACKNOWLEDGMENTS

Firstly, I would like to thank my principal supervisor Dr. Xin Yu, my co-supervisor Dr. Qian Peter Su, with my sincere and heartfelt gratitude and appreciation for the supervision journey that provided me with the guidance and counsel I needed to succeed in my Ph.D. program. They are great mentors in mapping my Ph.D. journey, advising on a research topic, and connecting me with the resources I need.

I would also like to thank my other mentors, collaborators and colleagues at University of Technology Sydney. I would like to thank Porf. Yi Yang, Prof. Yunchao Wei, Dr. Linchao Zhu, Dr. Ping Liu, Dr. Yu Wu, Dr. Jiaoxu Miao, Dr. Yawei Luo, Dr. Xiaohan Wang, Dr.Fan Ma, Dr.Xuanyi Dong, Dr. Yanbin Liu and many others. I was really fortunate to work with them and participate in intellectual conversations with them.

Lastly, I would like to thanks my mother, Qiufen Li, my father, Hongru Li, and my wife, Qianyu Feng, for their support and love throughout my Ph.D. journey.

Abstract

Despite the empirical and preliminary successes in computer vision, deep neural networks often require large-scale annotated training datasets. When applied to complex visual understanding problems in the real world, their performance is limited, since both data and annotations can be notoriously costly to collect, or may exist in various noisy or imperfect forms. Further, data annotating in such applications is also tedious to scale up, which demands highly skilled professionals, introducing challenges to use the cost-effective solutions, *e.g.*, crowdsourcing. Even worse, additional annotated data is always desired when the trained models need to be accordingly adapted to the dynamically changing environments. Thus, both the academic and industrial communities are calling for data-efficient deep learning algorithms.

In this thesis, we address the grand challenge of data-efficient and label-efficient visual understanding in realistic and imperfect real-world environments. To address this issue, we investigate deep learning approaches to leverage low-quantity training data and low-quality imperfect annotations. We propose a comprehensive suite of state-of-the-art approaches to tackle the data-efficient visual understanding from three directions, including : (1) applying *low-shot learning paradigms* that are intrinsically data-efficient, *e.g.*, few-shot learning or zero-shot learning. (2) exploiting imperfect labeled data to enable *learning with noise*. (3) *transferring prior knowledge* from the data-abundant domain into the data-hungry one. To demonstrate the effectiveness and efficiency in representative computer vision applications, extensive experiments are conducted on several dense prediction tasks, e.g., human parsing, scene parsing, semantic segmentation, and face super-resolution.

List of Publications

Related to the Thesis :

1. **P. Li**, Y. Xu, Y. Wei, and Y. Yang, "Self-Correction for Human Parsing," in *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

2. **P. Li**, Y. Wei, and Y. Yang, "Consistent structural relation learning for zero-shot segmentation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

3. **P. Li**, Y. Wei, and Y. Yang, "Meta parsing networks: Towards generalized few-shot scene parsing with adaptive metric learning," in *ACM International Conference on Multimedia (MM)*, 2020.

4. **P. Li**, X. Yu, and Y. Yang, "Super-resolving cross-domain face miniatures by peeking at one-shot exemplar," in *IEEE International Conference on Computer Vision (ICCV)*, 2021.

Others :

5. **P. Li**, X. Dong, X. Yu, and Y. Yang, "When Humans Meet Machines: Towards Efficient Segmentation Networks," in *The British Machine Vision Conference (BMVC)*, 2020.

6. **P. Li**, P. Pan, P. Liu, M. Xu, and Y. Yang, "Hierarchical Temporal Modeling with Mutual Distance Matching for Video Based Person Re-Identification," in *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2020.

7. X. Pan, **P. Li**, Z. Yang, H. Zhou, C. Zhou, H. Yang, J. Zhou, and Y. Yang, "In-N-Out Generative Learning for Dense Unsupervised Video Segmentation," in *ACM International Conference on Multimedia (MM)*, 2022.

8. Z. Yang, **P. Li**, Q. Feng, Y. Wei, and Y. Yang, "Going deeper into embedding learning for video object segmentation," in *IEEE International Conference on Computer Vision Workshops*, 2019.

9. Q. Feng, Z. Yang, **P. Li**, Y. Wei, and Y. Yang, "Dual embedding learning for video instance segmentation," in *IEEE International Conference on Computer Vision Workshops*, 2019.

10. X. Pan, H. Luo, W. Jiang, J. Zhang, J. Gu, and **P. Li**, "SFGN: Representing the Sequence with One Super Frame for Video Person Re-identification," in *Knowledge-Based Systems*, 2022.

TABLE OF CONTENTS

List of Figures

Figure **Page**

List of Tables

CHAPTER 1

INTRODUCTION

1.1 Research Background

Building machine intelligence that learns, thinks, and behaves like human beings is one of the main goals of deep learning and artificial intelligence. In recent years, the emergence of large-scale deep learning methods has resulted in critical breakthroughs in a wide range of computer vision tasks, including classification [4, 5], detection [6], segmentation [7, 8], *etc.*. For instance, the deep neural networks ResNet [4, 5] and ViT [9] achieve better classification accuracy than humans on ImageNet. While the increased learning capacities of the novel network architectures are only partly responsible for the advancements, the massive amount of training data also plays a significant role. Particularly, the success of these leading deep learning approaches highly relies on the abundant large-scale labeled datasets, yet their performance may inevitably degrade with imperfect data annotations or with less training data. In addition, when tackling more complex and challenging visual understating tasks, *e.g.*, dense prediction tasks, a massive quantity of high-quality data is required to achieve high-level performance.

However, in the real-world application, the large-quantity data is notoriously cost-intensive to collect and time-consuming to annotate. Moreover, the labeled data may exist in a variety of noisy and imperfect forms. Particularly, these kinds of data-inefficiency, in terms of both *data quantity* and *data quality*, usually happen in the complex dense prediction tasks, where more challenging pixel-level annotations are involved during the training process. To overcome the data-hungry issues in such vision applications, scaling

up the size of the annotated dataset is a painstaking process. The pixel-level image labeling may demand highly skilled annotators and limit the usage of cost-effective solutions, *e.g.*, crowdsourcing. Besides, owing to proprietary or confidential reasons, the crowdsourcing annotation of a large-scale dataset is usually infeasible. Furthermore, in real-world dynamically changing environments, additional annotated data is always desired when the trained models need to be accordingly adapted. Including but not limited to the aforementioned problems, it has sparked a heated debate calling for data-efficient deep neural networks in both the academic and industrial communities.

1.2 Data-efficient Visual Understanding

Learning toward data-efficient deep neural networks with low-quality data or with a small quantity of data is highly required. Consequently, this thesis investigates the data efficiency issue of deep learning approaches especially targeted at the dense prediction tasks. In this thesis, we mainly present data-efficient visual understanding algorithms to leverage (i) **low-quality** imperfect data to enable learning with noise and (ii) **low-quantity** labeled data by exploiting few-shot or zero-shot learning.

We show that deep learning methods for dense prediction tasks, *e.g.*, human parsing, scene parsing, semantic segmentation, and image generation, can still perform well with limited and imperfect data, which are highly practical and desired in real-world applications. To achieve better data effectiveness and efficiency, we focus on data-efficient approaches from three main directions by (i) applying *low-shot learning paradigms* that are intrinsically data-efficient, *e.g.*, few-shot learning or zero-shot learning, by (ii) exploiting imperfect labeled data to enable *learning with noise*, by (iii) *transferring prior knowledge* from the data-abundant domain into data-hungry one.

1.2.1 Low-shot Learning Paradigms for Data Efficiency

Conventional fully-supervised approaches require a large-scale dataset to train the deep learning models, and the quantity of the training data available has a significant impact on the performance of these models. However, in real-world applications, data annotation is a time-consuming and cost-intensive process that demands a significant amount of effort for each new task of interest. Exploiting other non-fully supervised learning paradigms would be one straightforward strategy to eliminate this data-dependency requirement. Such low-shot learning paradigms either require only a limited number

of data samples with supervised information (*i.e.* few-shot learning), or require zero training samples by transferring the prior knowledge from language modal to vision modal (*i.e.* zero-shot learning).

Few-shot learning (FSL) aims to generalize the learned knowledge to novel categories given only a few labeled training samples. Many meta-learning-based approaches have been proposed to address the few-shot problems, which can be roughly divided into gradient-based approaches [10] and metric-based approaches [11–14]. To be specific, the gradient-based methods search for a model weight configuration, which can be fast adapted to a novel task with only a few gradient update steps. Despite the competitive performance, the gradient-based approaches suffer from the need to perform additional fine-tune steps on a novel task. The metric-based approaches alternatively learn an embedding that can be used to perform the comparison between the labeled support samples and the unlabeled query ones. The recent progress on few-shot learning mainly focuses on addressing the image classification, while the more challenging dense prediction tasks, *e.g.*, semantic segmentation, are actually not well explored. In Chapter 4, we exploit the few-shot learning paradigm on the challenging scene parsing task. While, in Chapter 6, we investigate few-shot domain adaptation on the face super-resolution task.

Zero-shot Learning (ZSL) aims to recognize unseen classes with no training examples by leveraging the semantic label embeddings (*e.g.*, word embeddings or attribute vectors) as side information [15, 16]. Despite on the traditional image classification task, ZSL has been applied to predict novel action in videos [17, 18], detect unseen objects [19, 20], and recently, to segment pixel-wise unseen categories [21, 22]. Learning the project between semantic space and visual space is the key challenge. Former practices address ZSL by learning a projection function from visual space to semantic space [23, 24] or model weight space [25]. However, the intra-class variation in visual space is neglected by mapping to a deterministic word embedding in semantic space. Recently, due to the advance of deep generative models [26, 27], one can overcome the scarcity of unseen visual features by directly generating samples from semantic word embeddings. Commonly, these generative-based methods [28, 29] train their models firstly on seen classes and then generate visual features of unseen classes. Although the issue of data efficiency is partly addressed, the performance of the generative process solely relies on the generalization ability of the generator. In Chapter 5, we propose a structural constraint for the generative-based zero-shot learning paradigm to solve the semantic segmentation task.

1.2.2 Data-efficient Learning from Imperfect Annotations

Besides alleviating the quantity requirements to achieve data efficiency, exploiting low-quality data examples via learning with noise is one of the most widely considered directions. When annotating the dataset for deep learning algorithms, some non-expert methods like crowdsourcing have been widely applied to reduce the high labeling cost. However, annotating the pixel-level labels for dense prediction tasks could be challenging even for skilled domain specialists. Thus, unreliable annotations, *i.e.*, noisy or imperfect labels inevitably exist in real-world collected datasets. Training deep neural networks with the presence of these noisy labels may lead to poor performance and generalization ability since DNNs tend to overfit these corrupted labels. Hence, the critical challenge for data efficiency is achieving strong generalization capability even learning with the low-quality noise labels.

A line of studies has investigated the problem of learning with noise. Beyond the traditional machine learning techniques [30, 31], learning with noisy labels in deep learning has gained broad attention recently. Most recent efforts in learning with noise via deep neural networks either develop the robust architecture [32, 33] to process diverse label noises reliably or adjust the noise labels and loss values by sample selection strategies [34, 35]. For example, [36] averages model weights as self-ensembling and applies it in the semi-supervised learning task. In [37], they average the model weight and lead to better generalization. Meanwhile, pseudo-labeling [38, 39] is another typical technique used to assign pseudo-labels to correct the noise labels. Label smoothing [40] also addresses the problem of over-fitting and over-confidence by regularizing the one-hot label. Inspired by above works, in Chapter 3, we propose a general data-efficient framework to correct the noisy training labels via the model and label mutually promoting process. To the best of our knowledge, we have made the first attempt to formulate the label noise problem as a mutual model and label optimization in the fine-grained human parsing task.

1.2.3 Data Efficiency via Prior Knowledge Transferring

Once there is a large amount of labeled data, deep neural networks thrive at the visual understanding tasks, yet the performance may degrade severely when only limited supervision is available. On the other hand, humans and animals can learn about the novel classes of images in a data-efficient way, with only a few examples. A hypothesis is that intelligent systems benefit from structured *a priori* to achieve data efficiency. Among

several possible promising directions, we can implicitly or explicitly improve the deep learning models' performance and data efficiency by transferring the prior knowledge from the data-abundant domain to the data-hungry domain. Such prior knowledge could be inductive biases either from another domain of the current task or external world prior knowledge from a different modal.

Cross-Domain Inductive Biases as Prior Knowledge Domain adaptation is broadly studied in computer vision which deals with situations when a model trained on one source distribution is applied to a different but similar target distribution. In Chapter 6, with the prior knowledge in the data-abundant source domain, we bridge the domain gap and solve the cross-domain face super-resolution task in an extreme data-efficient way. To overcome the need of large-scale training data and improve the adaption ability of models on new domains, many works have been extensively proposed [10, 41–49]. Early one/few-shot-based classification tasks [50] construct generative models from shared appearance priors across classes for classification. Recently, a new stream of works focuses on using meta-learning to quickly adapt models to novel tasks [10, 51, 52]. However, these one/few-shot methods are mainly applied to different classification tasks without considering domain gaps between image pairs. Pix2Pix [53] and CycleGAN [54] have been proposed as image-to-image translation networks. However, due to the scarcity of samples in the target domain, these methods might not be suitable for transferring from the source domain to the target one with few samples. Motivated by these findings, in Chapter 6, we present a face super-resolution method that incorporates latent prior knowledge to smoothly adapt to the new target domain with only a one-shot exemplar.

Cross-Modal Inductive Biases as Prior Knowledge Beyond the prior knowledge transferring across different scenarios with a domain gap (*e.g.* domains with different lighting conditions), we further study the problem of cross-modal knowledge transferring. In Chapter 5, we present a zero-shot semantic segmentation method to transfer the cross-modal prior knowledge from language modal to vision modal. Semantic segmentation under fully supervised paradigm [48, 55–59] and domain adaptation scheme [60–62] are extensively studied. To extremely reduce the cost of label annotation, previous works focus on weakly-supervised segmentation [63–65] and few-shot segmentation [66, 67]. Most recent works [21, 22] further extend the zero-shot learning to the semantic segmentation task. Under such a setting, the semantic word embeddings are projected to synthetic visual features [22] and classifier weights [21]. In contrast, instead of simple node-to-node mapping, we tackle the zero-shot segmentation from a new perspective as structural relation learning from semantic space to visual space. In Chapter 5, the

semantic segmentation task is addressed in an extreme data-efficient way with zero training examples.

1.3 Thesis Organization

This thesis is organized as follows,

- *Chapter2*: This chapter presents the literature review on data-efficient machine learning techniques and covers related works on several dense prediction tasks introduced in this thesis, e.g., semantic segmentation, scene/human parsing and face super-resolution.

- *Chapter3*: This chapter addresses the data-efficient visual understanding by learning with noise and investigates the human parsing task with low annotation quality. We propose a simple yet effective, generic, model-agnostic framework called SCHP for human parsing tasks. We tackle the problem of human parsing tasks under learning with noise scenarios, which is never explored before. From this novel perspective, we unravel the problem by dealing with the pixel-level label noise during training process by self-correction mechanism. More specifically, we progressively promote the reliability of the supervised labels as well as the learned models. Our SCHP is model-agnostic and can be applied to any human parsing model to further enhance its performance. We achieve the new state-of-the-art results on six benchmarks, including LIP, Pascal-Person-Part, and ATR for single human parsing, CIHP and MHP for multi-person human parsing, and VIP for video human parsing tasks. This work was published on IEEE Transactions on Pattern Analysis and Machine Intelligence [58] and is also the winner solution for CVPR'19 Look Into Person Challenge (LIP) challenge.

- *Chapter4*: This chapter studies the data-efficient visual understanding from the few-shot learning perspective and investigates the few-shot segmentation task with low-quantity annotated training data. We advance the few-shot segmentation paradigm towards a more challenging yet general scenario, *i.e.*, Generalized Few-shot Scene Parsing (GFSP). In this task, we take a fully annotated image as guidance to segment all pixels in a query image. Our mission is to study a generalizable and robust segmentation network from the meta-learning perspective so that both seen and unseen categories can be correctly recognized. Accordingly, we present Meta Parsing Networks (MPNet) to better exploit the guidance information in

the support set. We conduct experiments on two newly constructed benchmarks, *i.e.*, *GFSP-Cityscapes* and *GFSP-Pascal-Context*. Extensive ablation studies well demonstrate the effectiveness and generalization ability of our MPNet. This work was published on ACM Multimedia Conference [68] as oral presentation.

- *Chapter5*: In this chapter, we explore the data-efficient visual understanding by transferring the cross-modal prior knowledge and investigates the zero-shot semantic segmentation task zero-shot training samples. Zero-shot semantic segmentation aims to recognize the semantics of pixels from unseen categories with zero training samples. We propose a Consistent Structural Relation Learning (CSRL) approach to constrain the generation of unseen visual features by exploiting the structural relations between seen and unseen categories. We observe that different categories usually have similar relations in either semantic word embedding space or visual feature space. This observation motivates us to harness the similarity of category-level relations on the semantic word embedding space to learn a better visual feature generator. We conduct extensive experiments on Pascal-VOC and Pascal-Context benchmarks. The proposed CSRL outperforms existing state-of-the-art methods by a large margin, resulting in ~7-12% on Pascal-VOC and ~2-5% on Pascal-Context. This work was published on Neural Information Processing Systems [69] as spotlight presentation.

- *Chapter6*: In this chapter, we explore the data-efficient visual understanding by transferring the cross-domain prior knowledge and propose a data-efficient approach for the face super-resolution (FSR) task under a low-quantity data scenario. Conventional face super-resolution methods usually assume that testing low-resolution (LR) images lie in the same domain as the training ones. Due to different lighting conditions and imaging hardware, domain gaps between training and testing images inevitably occur in many real-world scenarios. Neglecting those domain gaps would lead to inferior face super-resolution performance. However, how to transfer a trained FSR model to a target domain efficiently and effectively has not been investigated. To tackle this problem, we developed a Domain-Aware Pyramid-based Face Super-Resolution network named DAP-FSR network. Our DAP-FSR makes the first attempt to super resolve LR faces from a target domain by exploiting only a pair of high-resolution (HR) and LR exemplars in the target domain. Extensive experiments on three benchmarks validate the effectiveness and superior performance of our DAP-FSR compared to the state-of-the-art methods.

This work was published on The International Conference on Computer Vision [70].

- *Chapter7*: This chapter summarizes the thesis and shows directions for potential future improvements.

CHAPTER 2

LITERATURE REVIEW

In this thesis, we propose deep neural networks for several typical dense prediction tasks in computer vision, aiming to develop data-efficient deep learning algorithms. We try to answer the following questions, (i) what does it mean for machine intelligence to learn as quickly as human beings? (ii) what efforts should be made to relieve data hungriness in deep learning approaches? and (iii) what are the potential research directions to pursue? In this section, we list open problems and research directions from the most related works in two main aspects. Firstly, we review the related works for the visual understanding tasks introduced in this thesis. Then, we discuss several promising machine learning techniques to achieve data-efficient learning related to the content of this thesis.

2.1 Existing Data-hungry Visual Understanding Methods

Semantic Segmentation models [55, 71–75] target on performing pixel-wise classification for a given image, which mainly includes two basic tasks, *i.e.*, object semantic segmentation [76, 77] and scene parsing [78–82]. In particular, object semantic segmentation [48, 83–85] only considers to recognize the object of interest in the image while overlook the complex stuff semantics such as person [86, 87], car and road. Compared to object semantic segmentation, scene parsing [59, 88] poses a much general, practical yet challenging task, which requires all the pixels belonging to the given image to be

well classified. Although significant progress has been made due to the development of deep learning, both two tasks require all categories (seen) should be well defined before training, leading to the learned knowledge can not being transferred to recognize the new emerging categories (unseen). Additionally, annotating pixel-level labels is often costly in terms of both human efforts and finance, making the current state-of-the-art segmentation models not suitable for addressing generalized few-shot semantic segmentation problems as introduced in Chapter 4.

Human Parsing [89–91], as a fine-grained semantic segmentation task [59, 68, 69], has received more and more attention due to the potential application in human analysis, virtual reality, image editing *etc*. Several different aspects of human parsing tasks have been studied. Some early works [91–93] utilize pose estimation together with the human parsing simultaneously as a multi-task learning problem. In [94], they cooperate the edge prediction with human parsing to accurately predict the boundary area. Moreover, [95–98] study the human parsing task in a multi-person scenario, where not only to label the semantic parts but also distinguish human instances. Recent works [99, 100] further extend the image-based human parsing into a video-based application. Most of the prior works assume the fact that ground-truth labels are accurate and well-annotated. However, due to time and cost limitations, there inevitably exists lots of different label noises (as shown in Fig. 3.1). Meanwhile, it is impracticable to correct the pixel-wise labels manually. Guided by this intuition, we try to tackle this problem via a simple yet effective self-correction mechanism in Chapter 3.

Few-shot object segmentation [101–103] has received much attention recently due to its advantages in learning novel categories [47, 49] without much annotations. Most previous approaches [104, 105] follow the metric-based few-shot learning scheme and make great efforts to develop robust feature embedding to measure the pixel-wise similarity between the object from the support image and the query one. However, the current few-shot segmentation [66, 106, 107] only considers a simple case, *i.e.*, segmenting one or two objects from unseen categories in the given query image, which usually does not work well for the real scenario where pixels from dozens of unseen categories appear. Chapter 4 in this thesis takes one step further by extending the few-shot segmentation to a much more complex yet practical scene parsing problem. Unlike the object segmentation paradigm, the learned meta parser should be robust to multiple seen and unseen categories, simultaneously. To this end, we explore multiple strategies to endow the meta parser with better generalization ability. We compare the proposed solution with the state-of-the-art few-shot object segmentation approach [66],

and significant improvement is observed.

Generalized Zero-shot Semantic Segmentation Semantic segmentation under fully supervised paradigm [48, 55–59] and domain adaptation scheme [60–62] are extensively studied. To extremely reduce the cost of label annotation, previous works focus on weakly-supervised segmentation [63–65] and few-shot segmentation [66, 67, 108]. Most recent works [21, 22] further extend the zero-shot learning to the semantic segmentation task. The semantic word embeddings are projected to synthetic visual features [22] and classifier weights [21]. However, the structural relations between seen and unseen classes are not well explored. In Chapter 5, instead of simple node-to-node mapping, we tackle the zero-shot segmentation from a new perspective as structural relation learning from semantic space to visual space.

Face Super-Resolution (FSR) also known as face hallucination, aims at establishing the intensity relationships between input LR and output HR face images from the same domain. Traditional holistic appearance-based methods firstly leverage a parameterized model to represent faces and then construct the mappings between LR and HR faces. Some representative models super-resolve HR faces from LR ones by adopting global linear mapping [109, 110], or optimal transport [111]. However, they require input LR images aligned to a canonical pose and HR faces in the database to share similar facial expressions. Later on, part-based approaches have been proposed to relax the strict requirements in holistic appearance-based methods. Part-based face hallucination algorithms [112–114] firstly extract local facial regions and then upsample them separately.

Taking advantage of the powerful feature representation of deep neural networks, deep learning based face super-resolution methods [115, 115–123] have been proposed and achieved promising results. Several methods exploit prior knowledge, such as facial attributes [124], parsing maps [121], facial landmarks [125–127] and identity [128, 129], to advance the upsampling performance. However, when LR faces are captured from another domain, such as different imaging conditions, existing methods may fail to super-resolve them photo-realistically. Moreover, when the new domain data is not abundantly available, it would be challenging to retrain FSR networks with such a limited number of samples. Simple fine-tuning FSR networks with few samples does not solve this problem either. Therefore, previous methods may fail to authentically super-resolve LR faces existing a domain gap from the training domain when there are only one or few samples available from the new domain. In Chapter 6, we make the first attempt to address this

challenging scenario in a data-efficient manner.

2.2 Relevant Data-efficient Techniques

Pseudo-Labeling [38, 39] is a typical technique used in semi-supervised learning. In the semi-supervised learning setting, they assign pseudo-labels to the unlabeled data. However, in our task setting in Chapter 3, we are unable to locate the label noise since all ground truth is treated equally. Besides, from the perspective of distillation, the generated soft pseudo label contains many so-called *dark knowledge* [130] which could serve as the purification signal. And label smoothing [40] also addresses the problem of over-fitting and over-confidence by regularizing the one-hot label. Inspired by these findings, we design a cyclically learning scheduler to infer more reliable pseudo-masks by iteratively aggregating the current learned model with the former optimal one in an online manner. Furthermore, those corrected labels can, in turn, boost the model performance simultaneously.

Self-Ensembling There are a line of researches [36, 37, 131] that exploit self-ensembling methods in various scenarios. For example, [36] averages model weights as self-ensembling in the semi-supervised learning task. In [37], they average the model weight and lead to better generalization. Different from them, in Chapter 3, our proposed self-correction approach is to correct the noisy training label via a model and label mutually promoting process. In an online manner, we average both model weights and predictions simultaneously. To the best of our knowledge, we have made the first attempt to formulate the label noise problem as a mutual model and label optimization in fine-grained semantic segmenting to boost the performance. Furthermore, our proposed method is online learning with a cyclical scheduler and only exhaust little extra computation.

Few-shot learning aims to generalize the learned knowledge to novel categories with only a few labeled training samples. Many meta-learning-based approaches have been proposed to address the few-shot problems, which can be roughly divided into gradient-based approaches [10] and metric-based approaches [11–14, 132]. Specifically, the gradient-based methods search for a model weight configuration, which can be fast adapted to a novel task with only a few gradient update steps. Despite the competitive performance, the gradient-based approaches suffer from the need to perform additional fine-tune steps on a novel task. The metric-based approaches alternatively learn an embedding that can be used to compare the labeled support samples and the unlabeled query ones. The recent progress on few-shot learning mainly addresses the image classi-

fication task, while the much more challenging semantic segmentation task is not well explored. In Chapter 4 and Chapter 6, we investigate the generalized few-shot semantic segmentation and face super-resolution tasks from a few-shot learning perspective, respectively.

Zero-Shot Learning aims to recognize unseen classes with no training examples by leveraging the semantic label embeddings (*e.g.*, word embeddings or attribute vectors) as side information [15, 16]. Despite on the traditional image classification task, ZSL has been applied to predict novel actions in videos [17, 18, 133], detect unseen objects [19, 20], and recently, to segment pixel-wise unseen categories [21, 22]. Former practices address ZSL by learning a projection function from visual space to semantic space [23, 24] or model weight space [25]. However, the intra-class variation in visual space is neglected by mapping to a deterministic word embedding in semantic space. Recently, due to the advance of deep generative models [26, 27], one can overcome the scarcity of unseen visual features by directly generating samples from semantic word embeddings. Commonly, these generative-based methods [28, 29] train their models firstly on seen classes and then generate visual features for unseen classes. However, the quality of the generated unseen features solely relies on the generalization ability of the generator. Differently, in Chapter 5, we apply structural relation consistency as constraints to guide the learning process.

One-shot Domain Adaptation To overcome the need of large-scale training data and improve the adaption ability of models on new domains, many works have been extensively proposed [10, 41–49]. Early one/few-shot-based classification tasks [50] construct generative models from shared appearance priors across classes for classification. Recently, a new stream of works focuses on using meta-learning to quickly adapt models to novel tasks [10, 51, 52]. However, these one/few-shot methods are mainly applied to different classification tasks without considering domain gaps between image pairs.

Pix2Pix [53] and CycleGAN [54] have been proposed as image-to-image translation networks. However, due to the scarcity of samples in the target domain, these methods might not be suitable for transferring from the source domain to the target one with few samples. To mitigate the data hungry problem of deep neural networks, several works employ shared [41] or partially shared [134] latent space assumption to conduct image-to-image translation tasks, such as style transfer [41, 135] and face generation [136]. Since these methods only address the domain gap without learning the mapping between LR and HR images, they are unsuitable for face super-resolution. Instead, we investigate

the data-efficient one-shot domain adaptation problem in Chapter 6.

CHAPTER 3

SELF-CORRECTION FOR HUMAN PARSING

3.1 Preface

Human parsing, as a fine-grained semantic segmentation task, aims to assign each image pixel from the human body to a semantic category, *e.g.* arm, leg, dress, skirt. Understanding the detailed semantic parts of humans is crucial in several potential application scenarios, including image editing, human analysis, virtual try-on and virtual reality. Recent advances on fully convolutional neural networks [137, 138] have achieved various of well-performing methods for the human parsing tasks [91, 94].

To learn reliable models for human parsing, a large amount of pixel-level masks are required as supervision. However, labeling pixel-level annotations for human parsing is much harder than the traditional pixel-level understanding tasks. In particular, in the traditional semantic segmentation tasks [137, 138], all the pixels belonging to one instance share the same semantic label, which is usually easy to be identified by annotators. Differently, the human parsing task requires annotators to carefully distinguish detailed semantic parts of one person. Moreover, the situation will become even more challenging when the annotator got confused by the ambiguous boundaries between different semantic parts. Due to the aforementioned factors, there inevitably exist different types of label noises (as illustrated in Fig. 3.1) caused by the careless observations by annotators. This incomplete and low quality of the annotation labels will set a significant obstacle, which prevents the performance of human parsing from increasing to a higher level.

Figure 3.1: Different types of label noises in ground-truth annotations. The upper row shows the original images. The lower row shows the original ground-truth labels. Different types of noisy labels are illustrated from left to right, (a) coarse annotation around the boundary area; (b) confused fine-grained categories, where the upper-cloth is mislabeled as the coat; (c) confused mirror categories, where the right leg is mislabeled as the left leg; (d) multiple-person occlusion. Annotation noises are marked in white dashed boxes.

Previous efforts [90, 139–141] are all dedicated to design various frameworks to segment human parts and based on the assumption that all the annotated ground-truth masks are accurate. In this work, we tackle the human parsing task from a totally new perspective and investigate the problem of learning with inaccurate ground-truth masks. Our target is to improve the model performance and generalization ability by progressively refining the noisy labels during the training stage.

To this end, we introduce a noise-tolerant approach named Self-Correction for Human Parsing (SCHP), which can progressively promote the reliability of the supervised labels, as well as the learned models during the training process. Concretely, the whole SCHP pipeline can be divided into two sub-procedures, *i.e.*, model aggregation and label refinement. Starting from a model trained on inaccurate annotations as initialization, we design a cyclically learning scheduler to infer more reliable pseudo masks by iteratively aggregating the current learned model with the former sub-optimal one in an online manner. Besides, those corrected labels can in turn to boost the model performance, simultaneously. In this way, the self-correction mechanism will enable the learned models and the refined labels to mutually promote their counterpart, leading the future

models and labels to be more robust and accurate as the training goes on.

Our SCHP is a model-agnostic noise-tolerant strategy, which can be easily applied to different human parsing frameworks for further improving their performance. To validate the generalization ability of our SCHP, we conduct extensive experiments with four popular human parsing frameworks, including Deeplab V3+ [71], CE2P [94], OCR [56] and CE2P+ (a upgraded version CE2P), and consistent improvements can be observed on popular human parsing benchmarks, *i.e.*, LIP [91], PASCAL-Person-Part [142] and ART [139]. Particularly, we achieve the mIoU score of 59.36% on LIP, which outperforms the state of the art performance by more than 1.62%. Besides, our SCHP can be easily extended to address multiple human parsing and video human parsing problems. With the help of SCHP, we achieve the state-of-the-art performance of 45.25 (mAP^p), 51.08 (mAP^r), and 51.41 (mAP^r) on MHP v2.0 [96], CIHP [141], and VIP [100], respectively, outperforming other approaches by more than 2.55, 7.49, and 27.31. Moreover, together with advanced techniques and tricks, *e.g.*, multi-scale training and model ensembling, we rank the 1st place of all human parsing tracks (*i.e.*, 65.18% ($mIoU$) for single human parsing, 55.01% (mAP^r) for multiple human parsing and 52.97 (mAP^r)% for video human parsing) in the 3rd Look Into Person Challenge (in conjunction with CVPR 2019), which are 1.05%, 0.96% and 4.25% higher than the runner-up teams, respectively.

On the whole, our major contributions can be summarized as follows,

- We propose to tackle the challenging human parsing task by considering the label noises existing in the ground-truth masks. To the best of our knowledge, this is a new perspective in this research area, which is not well explored before.
- We propose a simple yet effective noise-tolerant approach named SCHP for alleviating the existing label noises, accordingly. By alternatively performing model aggregating and label refining in an online manner, SCHP could mutually promote the model performance and label accuracy.
- Our SCHP is model-agnostic, and thus can be applied to various human parsing frameworks. Extensive ablation experiments well demonstrate the generalization ability and the superiority of the proposed SCHP.
- Benefiting from the proposed SCHP, this work achieves new state of the art performance on six single/multiple human parsing benchmarks, and won the winner prize of all three human parsing tracks in the 3rd Look Into Person Challenge.

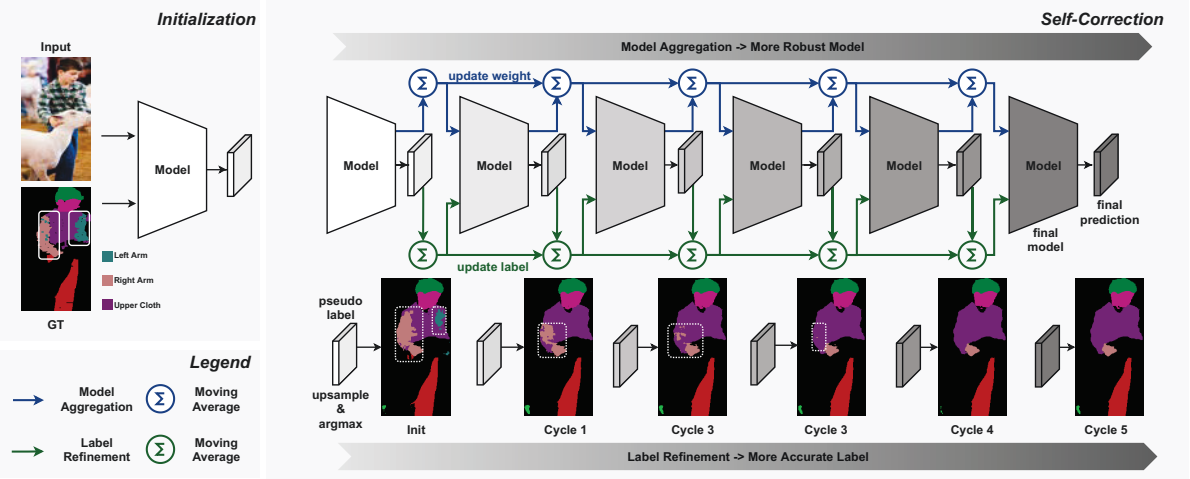


Figure 3.2: Overview of the SCHP pipeline. Starting from the warm-up initialization by training with inaccurate annotations, we design a cyclically learning scheduler to infer more reliable pseudo masks through iteratively aggregating the current learned model with the former optimal one in an online manner. Besides, those corrected labels can in turn to boost the model performance, simultaneously. In this way, the models and the masks get more robust and accurate during the self-correction cycles. Label noises are specially marked in white boxes.

3.2 Self-correction for Human Parsing

3.2.1 Overview

Given a training dataset with images $\mathscr{X}$ and inaccurate mask annotations $\mathscr{Y}$ with label noise, our method aims to train a noise-robust human parser, by simultaneously benefiting from model aggregation, label refinement and their interaction. We first warm-up the human parser by training the model using the original inaccurate annotations. After the warm-up initialization, we carry out an alternating optimization between models and mask annotations. We cyclically aggregate the learned model and refine the mask annotations $\mathscr{Y}$ in a moving average manner to build a noise-robust human parser. Concretely, within one self-correction cycle, we first learn the segmentation model using the mask annotations $\mathscr{Y}$ refined in the previous cycle. Then, a parameter-wise moving average operation is conducted to aggregate the learned model weights in the current cycle with the sub-optimal model from the previous cycle. Finally, we leverage the aggregated model to infer the mask predictions of all training images, which are further employed to correct the noise in annotations via a pixel-wise moving average operation. The refined mask annotations will serve as ground-truth to supervise the training for

the next cycle. Fig. 3.2 shows an overview of the proposed self-correction mechanism for human parsing. In the following, we give more details about our self-correction human parsing (SCHP) with model aggregation and label refinement procedure.

3.2.2 Single-person Human Parsing

Intuitively, when training a human parser model with the noisy annotations, the predictions for those falsely annotated regions often tend to be uncertain. For instance, if a region of *dress* is wrongly labeled as *skirt*, the learned knowledge from other correctly labeled *dress* regions will make the model produce inconsistent semantic prediction (*dress*) compared to its false ground-truth (*skirt*) during the training stage. Therefore, if we can successfully take advantage of the model predictions, the existing noises from ground-truth annotations will be partially surpassed, resulting in a better human parser accordingly. To this end, we propose a simple yet effective self-correction training approach, which includes three basic steps, *i.e.*, warm-up internalization, online model aggregation and online label refinement. As far as we know, it is the first attempt to tackle the human parsing task from the label noise perspective.

Warm-up Initialization The potential performance promotion relies on the initial performance of the model. In other words, if the intermediate results generated by the network are not accurate enough, they may potentially harm the following self-correction process. Therefore, we start to run our proposed self-correction algorithm after a good initialization, *i.e.*, when the training loss starts to flatten with the original ground-truth annotations. To avoid introducing an extra training cost, we shorten the initial warm-up stage and keep the same total training epochs to make a fair comparison with other methods.

Online Model Aggregation We aim to discover all the potential information from the past sub-optimal models to improve the performance of the future model. Intuitively, the learned models will converge to different local-minimums based on different initialized parameters at the beginning of training. There exists great model disparity among these sub-optimal models and assembling their complementary knowledge tends to produce a better model state. Motivated by this, we propose to successively perform multiple training cycles and progressively aggregate the learned model from each cycle in an online manner to generate the final one. Formally, we denote the set of the models obtained in different self-correction cycles as $\{\boldsymbol{\theta}_i\}_{i=1}^{N}$ and N is the total number of self-correction cycles. Suppose $\boldsymbol{\theta}$ be the model learned at the end of current cycle i. We then aggregate $\boldsymbol{\theta}$ with the former sub-optimal one $\boldsymbol{\theta}_{i-1}$ to output an updated model weight $\boldsymbol{\theta}_i$

as the initial model weight for the next self-correction cycle via a parameter-wise moving average operation,

$$\boldsymbol{\theta}_i = \frac{i}{i+1}\boldsymbol{\theta}_{i-1} + \frac{1}{i+1}\boldsymbol{\theta}. \tag{3.1}$$

However, we experimentally find that simply performing the parameter-wise moving average operation would lead to even worse performance, which is caused by the inaccurate parameter estimation (mean: μ, variance: σ^2) after model aggregation for BatchNorm [143] layers. To tackle this issue, we forward all the training samples for one epoch to exactly re-estimate the BatchNorm statistics in all BatchNorm layers as follows,

$$\begin{aligned} m &= (t-1)/t \\ \mu_t &= m\mu_{t-1} + (1-m)E[x_B] \\ \sigma_t^2 &= m\sigma_{t-1}^2 + (1-m)Var[x_B], \end{aligned} \tag{3.2}$$

where t is the iteration number in one epoch and x_B is the input features to the batch norm layer. Therefore, we need to apply additional computation overhead to perform model aggregation and re-estimate Batchnorm parameters at the end of each self-correction training cycle. However, since the model aggregation contains only simple average operation and the Batchnorm re-estimation can also be completed with one network forwarding epoch, the additional computation overhead from both memory and time can be totally negligible compared to the self-correction training. As a result of the proposed model aggregation, with the self-correction process goes on, the network leads to wider model optima as well as improves the model's generalization ability.

Online Label Refinement It is known that soft, multi-class labels contain more dark information [130] compared with the one-hot labels. We aim to additionally explore all this dark information to improve the model performance and alleviate the label noises beyond the online model aggregation. After updating the model weight as mentioned in Eq. equation 3.1, we also update the ground-truth of training labels. These generated pseudo-masks are more unambiguous, smooth and have the relational information among the fine-grain categories, which are taken as the supervised signal for the next cycle's optimization. During successive self-correction cycles, the learned knowledge from correct annotations will potentially alleviate or eliminate the incorrect ones in the original ground-truth. The amended information within the pseudo-masks will then help improve the robustness of the learned model.

Here we denote the predicted label set obtained in different training cycles as $\{\mathscr{Y}_i\}_{i=1}^{N}$ and N is the total number of self-correction cycles. Same as the model aggregation process, we refine the ground-truth label $\mathscr{Y}$ generated by current optimal model $\boldsymbol{\theta}_i$ with

Algorithm 1: Self-Correction for Human Parsing

Input: Warm-up initialized model weight $\boldsymbol{\theta}$, original ground-truth mask annotations $\mathscr{Y}$, epoch number of each self-correctin cycle T, number of self-correction cycles N

Output: Human parser model $\boldsymbol{\theta}_N$

Initialize the model weight $\boldsymbol{\theta}_0 \leftarrow \boldsymbol{\theta}$;

Initialize the pseudo-mask $\mathscr{Y}_0 \leftarrow \mathscr{Y}$;

for $i \leftarrow 1,2,...,N$ **do**

  for $T_{cur} \leftarrow 1,2,..T$ **do**

    Update the learning rate η by Eq. equation 3.4;

    for *each batch in training set* **do**

      Calculate loss $\mathscr{L}$ using refined $\mathscr{Y}_{i-1}$;

      Gradient descending $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla \mathscr{L}$;

    end

  end

  Model aggregation by equation 3.1 to update $\boldsymbol{\theta}_i$;

  Update model weight $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_i$;

  Re-calculate the BN layer parameters by equation 3.2 ;

  Re-calculate the pseudo-mask $\mathscr{Y}$ using $\boldsymbol{\theta}_i$;

  Label refinement by equation 3.3 to update $\mathscr{Y}_i$;

end

the former sub-optimal one $\mathscr{Y}_{i-1}$ to output an updated pseudo-mask $\mathscr{Y}_i$ as the initial mask annotations for the next self-correction cycle via a pixel-wise moving average operation as follows,

$$\mathscr{Y}_i = \frac{i}{i+1}\mathscr{Y}_{i-1} + \frac{1}{i+1}\mathscr{Y}. \tag{3.3}$$

Cyclical Training Strategy We design the self-correction training with two main principles. First, at the end of each self-correction cycle, the network should converge to an acceptable sub-optimal state, which implies having a small learning rate. Second, in order to ensure the network jumping out of the former local optima and having sufficient model disparity between two self-correction cycles, we need to have a large learning rate at the beginning of each self-correction cycle. Therefore, during the self-correction cycles, we apply the learning scheduler to a cyclically annealing one [144]. Suppose each cycle totally contains T training epochs, in practice, we use a cosine annealing learning rate scheduler with cyclical restart [145]. Formally, η_{max} and η_{min} are set to the initial learning rate and final learning rate, while T_{cur} is the number of epochs since the last restart. Thus, the overall learning rate can be formulated as,

$$\eta = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})(1 + \cos(\frac{T_{cur}}{T}\pi)). \tag{3.4}$$

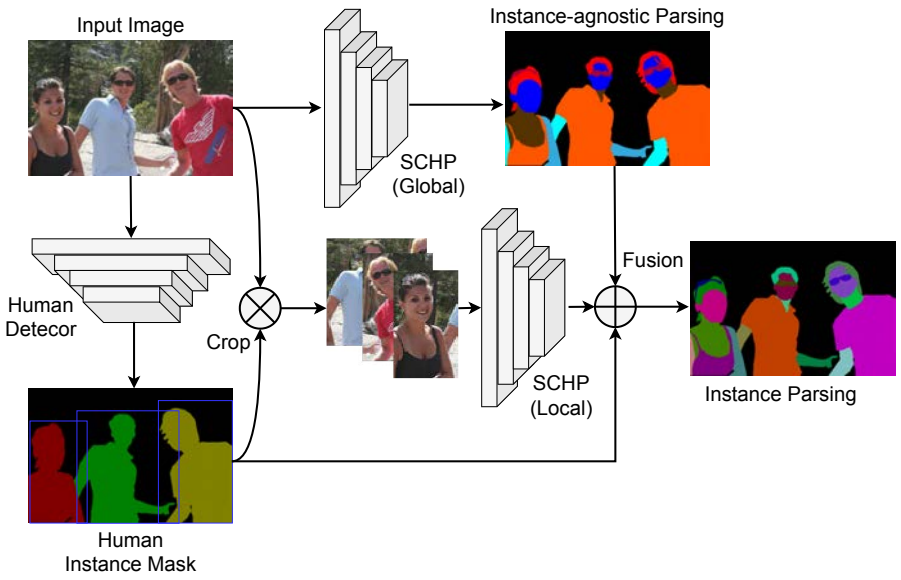


Figure 3.3: The pipeline for the multiple-person human parsing and video human parsing task.

Equipping with the cyclical annealing learning rate scheduler, the model aggregation and label refinement processes are mutually improving each other step-by-step after each cyclical training process. The proposed SCHP is training in an end-to-end online manner. The details of our proposed self-correction procedure are summarized in Algorithm 1.

3.2.3 Extension to Multi-person Human Parsing

Our SCHP is a general solution, which can be easily extended to tackle much more challenging multiple human parsing tasks. In this section, we give the details of two advanced applications of our SCHP, *i.e.*, multi-person human parsing and video multi-person human parsing.

Multi-person Human Parsing As a more challenging task, multiple-person human parsing aims to semantically categorize every pixel, as well as identify every human instance in the images. The main difference between the multiple-person parsing and the single-person parsing lies in distinguishing the human instances from each other. To maintain simplicity, a stage-wise training procedure is adopted to decouple this challenging task into human detection and single-person human parsing. Our framework is a top-down based pipeline, which is illustrated in Fig. 3.3.

An off-the-shelf advanced human detector (*i.e.*, Mask R-CNN [6]) is instantiated to served as a base detector to find human instances from the input images. We first

fine-tune this human detector to better predict human instances and keep it fixed during the training process of human parsing. Consequently, some candidate instances with their extended local contexts can be cropped from the original images. Then, based on the proposed SCHP model for single human parsing, a two-branch structure is constructed and trained from a general global view to a local view. The global SCHP branch is trained using the whole original images, which leverages most context information of images and learns the spatial relation information between several instances under the crowded scenario. The local SCHP branch focuses on more precisely parsing with a local context of an instance, which captures more detailed information. Both the global and local SCHP models have the same identical architecture but do not share the model weight.

The instance-agnostic parsing results (upper right of Fig. 3.3) can be obtained from the global branch. These instance-agnostic parsing results are further transformed into instance-aware parsing results, by fusing the human instance masks produced by the human detector with results from both global and local branches. To deal with the miss-matching issues between human instance masks and the instance-agnostic parsing results during the assignment, a breadth-first searching (BFS) label refinement post-processing is adopted following [94].

Video Multiple-person Human Parsing Video human parsing is the task for simultaneously identifying instances and recognizing multiple semantic parts of humans from video frames. We consider this task as the frame-based multiple human parsing task, which not only needs to recognize every human instance in one single frame but also identifies the human instance between different frames. First, we feed each frame from the video sequences into the above mentioned multiple-person human parsing framework to acquire human instance and parsing results. Then, we identify and link the human instances along the temporal dimension following DeepSORT algorithm [146].

3.3 Experiments: Single-person Human Parsing

Our SCHP is a model-agnostic mechanism and can be applied to any human parsing models for further enhancing their performance by alleviating the label noise under complex scenarios. In this section, we perform a comprehensive comparison of our SCHP with other single-person human parsing state-of-the-art methods, along with thorough ablation experiments to demonstrate the effectiveness of each component in SCHP.

Table 3.1: Comparisons on the LIP `validation` set. The symbol † marks the single-scale testing result.

Method	hat	hair	glove	s-glass	u-clot	dress	coat	sock	pant	j-suit	scarf	skirt	face	l-arm	r-arm	l-leg	r-leg	l-shoe	r-shoe	bkg	mIoU
Attention [147]	58.87	66.78	23.32	19.48	63.20	29.63	49.70	35.23	66.04	24.73	12.84	20.41	70.58	50.17	54.03	38.35	37.70	26.20	27.09	84.00	42.92
DeepLab [138]	59.76	66.22	28.76	23.91	64.95	33.68	52.86	37.67	68.05	26.15	17.44	25.23	70.00	50.42	53.89	39.36	38.27	26.95	28.36	84.09	44.80
SSL [90]	58.21	67.17	31.20	23.65	63.66	28.31	52.35	39.58	69.40	28.61	13.70	22.52	74.84	52.83	55.67	48.22	47.49	31.80	29.97	84.64	46.19
MMAN [148]	57.66	65.63	30.07	20.02	64.15	28.39	51.98	41.46	71.03	23.61	9.65	23.20	69.54	55.30	58.13	51.90	52.17	38.58	39.05	84.75	46.81
MuLA [93]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	49.30
JPPNet [91]	63.55	70.20	36.16	23.48	68.15	31.42	55.65	44.56	72.19	28.39	18.76	25.14	73.36	61.97	63.88	58.21	57.99	44.02	44.09	86.26	51.37
Deeplabv3+† [71]	65.02	71.06	37.98	31.37	68.33	28.25	54.14	47.67	73.99	27.12	17.92	25.00	73.99	63.64	65.69	57.35	56.99	43.97	44.74	87.51	52.09
CE2P† [94]	65.29	72.54	39.09	32.73	69.46	32.52	56.28	49.67	74.11	27.23	14.19	22.51	75.50	65.14	66.59	60.10	58.59	46.63	46.12	87.67	53.10
BraidNet [149]	88.0	66.8	72.0	42.5	32.1	69.8	33.7	57.4	49.0	74.9	32.4	19.3	27.2	74.9	65.5	67.9	60.2	59.6	47.4	47.9	54.4
OCR†[56]	67.42	73.16	44.01	36.31	69.90	35.53	57.03	51.87	75.09	29.85	19.16	26.63	75.95	66.81	68.65	63.23	62.12	50.60	50.68	87.96	55.60
CNIF [150]	69.55	73.45	45.17	41.45	**70.57**	38.52	57.94	54.02	75.07	28.00	**31.92**	30.20	76.38	68.28	69.49	65.52	65.51	52.67	53.38	87.99	57.74
SCHP(Ours)†	69.96	73.55	50.46	40.72	69.93	39.02	57.45	54.27	76.01	32.88	26.29	31.68	76.19	68.65	70.92	67.28	66.56	55.76	56.50	88.36	58.62
SCHP(Ours)	**70.63**	**74.09**	**51.40**	**41.70**	70.56	**40.06**	**58.17**	**55.17**	**76.57**	**33.78**	26.63	**32.83**	**76.63**	**69.33**	**71.76**	**67.93**	**67.42**	**56.56**	**57.55**	**88.40**	**59.36**

Table 3.2: Comparisons on the PASCAL-Person-Part `test` set. The symbol † marks the single-scale testing result.

Method	head	torso	u-arm	l-arm	u-leg	l-leg	bkg	mIoU
Attention [147]	81.47	59.06	44.15	42.50	38.28	35.62	93.65	56.39
HAZN [151]	80.76	60.50	45.65	43.11	41.21	37.74	93.78	57.54
LG-LSTM [89]	82.72	60.99	45.40	47.76	42.33	37.96	88.63	57.97
SS-JPPNet [91]	83.26	62.40	47.80	45.58	42.32	39.48	94.68	59.36
MMAN [148]	82.58	62.83	48.49	47.37	42.80	40.40	94.92	59.91
G-LSTM [140]	82.69	62.68	46.88	47.71	45.66	40.93	94.59	60.16
Part FCN [92]	85.50	67.87	54.72	54.30	48.25	44.76	95.32	64.39
PCNet [152]	86.81	69.06	55.35	55.27	50.21	48.54	96.07	65.90
Deeplab [138]	-	-	-	-	-	-	-	64.94
WSHP [153]	87.15	72.28	57.07	56.21	52.43	50.36	**97.72**	67.60
PGN [141]	**90.89**	**75.12**	55.83	64.61	55.42	41.57	95.33	68.40
CNIF [150]	88.02	72.91	64.31	63.52	55.61	54.96	96.02	70.76
SCHP(Ours)†	87.00	72.27	64.10	63.44	56.57	55.00	96.07	70.63
SCHP(Ours)	87.41	73.80	**64.98**	**64.70**	**57.43**	**55.62**	96.26	**71.46**

3.3.1 Experiment Settings

Datasets We evaluate our proposed method on three single-person human parsing benchmarks, including LIP [91], PASCAL-Person-Part [142] and ATR [139]. LIP [91] is the largest human parsing dataset, which contains 50,462 images with elaborated pixel-wise annotations with 19 semantic human part labels. The images collected from the real-world scenarios contain human appearing with challenging poses and views, heavily occlusions, various appearances and low-resolutions. LIP is divided into 30,462 images for `train` set, 10,000 images for `validation` set and 10,000 for `test` set. PASCAL-Person-Part [142] is a relatively small dataset annotated from PASCAL VOC 2010, including six semantic parts, *i.e.*, head, torso, upper/lower arms, upper/lower legs and one background class. It contains 1,716 and 1,817 images for `train` and `validation` sets, respectively. ATR [139] is another large-scale dataset that targets on fashion AI, which includes 18

Table 3.3: Comparison on the ATR `test` set.The symbol † marks the single-scale testing result.

Methods	pixel Acc.	F.G. Acc.	Precision	Recall	F1
ATR [154]	91.11	71.04	71.69	60.25	64.38
DeepLab [138]	94.42	82.93	78.48	69.24	73.53
PSPNet [7]	95.20	80.23	79.66	73.79	75.84
Attention [147]	95.41	85.71	81.30	73.55	77.23
DeepLabV3+† [71]	95.96	83.04	80.41	78.79	79.49
Co-CNN [139]	96.02	83.57	84.95	77.66	80.14
TGPNet [155]	**96.45**	87.91	83.36	80.22	81.76
CNIF [150]	96.26	87.91	84.62	**86.41**	85.51
SCHP(Ours)†	96.14	87.82	84.28	85.78	85.02
SCHP(Ours)	96.25	**87.97**	**84.99**	86.13	**85.55**

fine-grained semantic labels similar to LIP. The dataset contains 17,700 images which are split into 16,000 for training, 700 for validation and 1,000 for testing.

Evaluation Protocols Following common practice, we mainly report three standard metrics, including pixel accuracy (pixel acc), mean accuracy (mean acc), mean intersection over union (mIoU). The mIoU is the main metric to generally judge the overall parsing performance of the method. For ATR dataset, we report the average precision, recall and F1-score in order to make a fair comparison with previous works.

Implementation Details We choose the ResNet-101 [5] as the backbone of the feature extractor and use an ImageNet [156] pre-trained weights. And all compared methods adopt the same backbone model for fair comparisons. Specifically, we fix the first three residual layers and set the stride of the last residual layer to 1 with a dilation rate of 2. In this way, the final output is enlarged to 1/16 resolution size *w.r.t* the original image. We use 473×473 as the input resolution. Training is done with a total batch size of 32. The initial learning rate is set as 7e-3. We train our network for 150 epochs in total for a fair comparison, the first 100 epochs as initialization following 5 cycles each contains 10 epochs of the self-correction process. During testing, unless otherwise motioned, following general protocol [94, 150], we average the per-pixel classification scores at multiple scales with flipping, *i.e.*, the scale is 0.5 to 1.5 (in increments of 0.25) times the original size.

3.3.2 Model-agnostic Study

We first validate the model-agnostic characteristic of our SCHP in collaborating with any single person human-parsing frameworks. To this end, we choose four state-of-the-art frameworks as strong baselines to illustrate the effectiveness of our SCHP, which are briefly described below.

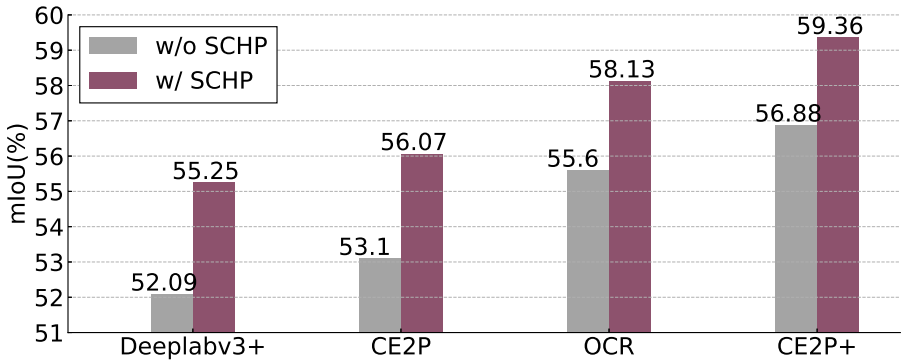


Figure 3.4: Model-agnostic study. The mIoU performance with different state-of-the-art models on LIP `val` set.

- **DeepLabV3+** [71] is one of the most popular state-of-the-art models for the semantic segmentation, which adopts an effective decoder module with atrous separable convolution to refine the segmentation results. It contains rich semantic information from the encoder module by applying spatial pyramid pooling, while the detailed semantic segmentation results are recovered by the simple yet effective decoder module.

- **CE2P** [94] is the winner solution for the human parsing tasks of the 2nd LIP challenge, which cooperates the edge prediction with human parsing to accurately predict the boundary area. CE2P employs an edge prediction branch to generate boundary-aware feature embedding, which is further concatenated with the semantic-aware feature embedding to produce a refined human parsing prediction.

- **OCR** [56] is a competitive semantic segmentation model that focuses on the context aggregation strategy. It utilizes the object-contextual representations, characterizing a pixel by exploiting the representation of the corresponding object class, achieving better performance than CE2P.

- **CE2P+** is an extension of CE2P, which is firstly proposed in this work. Based on CE2P, we make the following modifications. First, we additionally introduce a tractable surrogate loss function for optimizing the mIoU directly followed by [157]. Second, we introduce a regularization term by explicitly maintaining the consistency between the parsing prediction and the boundary prediction.

We show the comparisons of 'w/' and 'w/o' self-correction mechanism in Fig. 3.4. It can be observed that our SCHP is indeed a generic and model-agnostic approach, which brings

Figure 3.5: Visualization of SCHP results on LIP `val` set. The first row shows the original input images. The middle row shows the ground-truth labels. Different human categories are shown in colors in the third row.

consistent performance gain regardless of the model itself. Particularly, by incorporating with SCHP, DeeplabV3+, CE2P, OCR and CE2P+ have a mIoU performance improvement of +3.16, + 2.97, +2.43 and +2.48, respectively. *In the following*, we denote the term SCHP as the framework based on our CE2P+ model for comparison, and conducting all ablation experiments using CE2P+ to verify the effectiveness of each proposed component of SCHP.

3.3.3 Comparison with the State-of-the-art Approaches

We first compare the performance of our SCHP with other state-of-the-art methods on LIP in Table 3.1. It can be observed that the proposed SCHP outperforms all the other state-of-the-art methods, which well demonstrates its effectiveness. Particularly, our SCHP outperforms the current state-of-the-art model [150] by a large margin of 1.62%, which is a significant improvement considering the performance at this level. In addition, our SCHP achieves large gains especially for some categories with less pixel-level annotations like *scarf, sunglasses* and some confusing categories such as *dress, skirt* and left-right confusion. The gains are mainly from using both model aggregation and label refinement during the self-correction process. Furthermore, the qualitative comparison between the predicted results of SCHP and ground-truth annotations is shown in Fig. 3.5. We can observe that our SCHP can achieve even better parsing results than the original ground-truth ones for some images.

To validate the generalization ability of our method, we further report the comparisons on PASCAL-Person-Part dataset in Table 3.2 and on ATR dataset in Table 3.3.

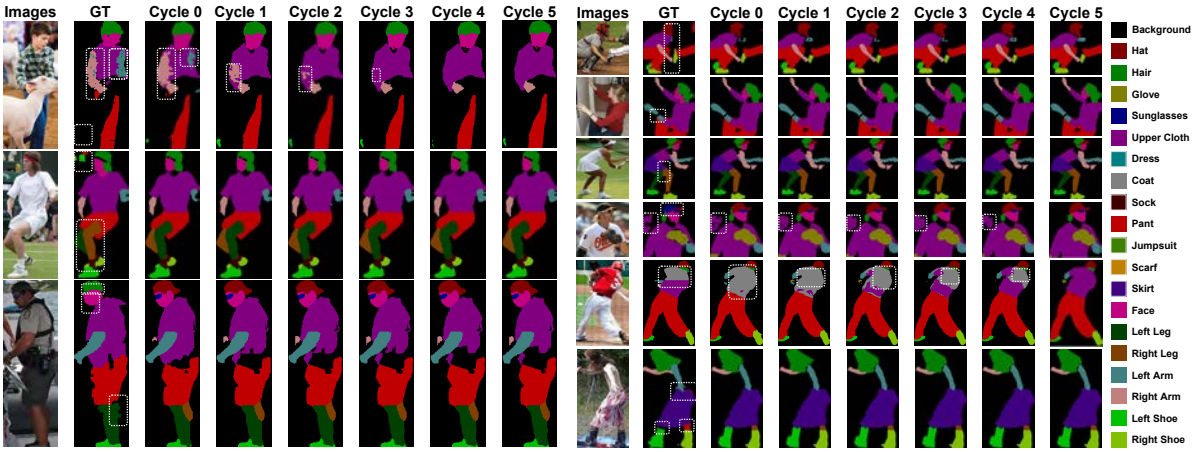


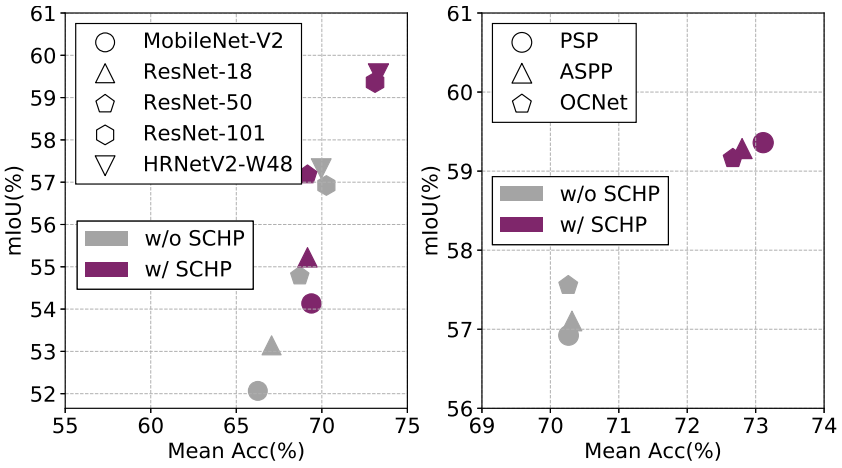
Figure 3.6: Examples from LIP `train` set during our self-correction process. Label noises like inaccurate boundary, confused fine-grained categories, confused mirror categories, multiple person occlusion are alleviated and resolved during the process. The boundaries of our corrected label are prone to be more smooth than the ground-truth label. Label noises are highlighted by white dotted boxes. Better zoom in to see the details.

It can be observed that our SCHP outperforms all the previous approaches. All these results well demonstrate the superiority and generalization of the proposed SCHP.

3.3.4 Ablation Experiments

We perform extensive ablation experiments to analyze the robustness of SCHP against different modules and the effect of each component in our SCHP. All experiments are conducted on LIP benchmark.

The Robustness against Different Modules Except for alternating entire models, we could also plug-and-play with various backbones and context encoding modules using the CE2P+ framework. Fig. 3.7a shows SHCP with different backbones from lightweight model MobileNet-V2 [158] to relatively heavy backbone HRNetV2-W48 [159]. It is noteworthy that the lightweight MobileNet-V2 achieves the mIoU score of 52.1, which can be further enhanced to 54.1 benefiting from SCHP. This result is even better than some previous results [94] achieved by ResNet-101. We note that deeper network (18 *vs.* 50 *vs.* 101) tends to perform better. Regardless of different backbones, our SCHP brings consistent gains of 2.1, 1.7 and 2.5 in terms of mIoU, respectively. Besides, we further examine the robustness of our SCHP by varying the context encoding module, as shown in Fig. 3.7b. We choose three different types of modules, including multi-level global average pooling based module pyramid scene parsing network (PSP) [7],



(a) Backbone Module. (b) Context Encoding Module.

Figure 3.7: Robustness of SCHP against (a) different backbones and (b) context encoding modules. Experiments are conducted on LIP `val` set.

multi-level atrous convolutional pooling based module atrous spatial pyramid pooling (ASPP) [8] and attention-mechanism based module OCNet [55]. Despite the similar basic performance of these three modules, our SCHP obtains mIoU gains of 2.5, 2.1, 1.7 for PSP, ASPP and OCNet, respectively. This further highlights the robustness of the self-correction mechanism against different advanced segmentation modules.

Effect of Self-Correction In Table 3.4, we validate the effect of each component in our SCHP, including the model aggregation (MA) process and the label refinement (LR) process. When there are no MA and LR involved, our method degenerates to the conventional training process. We can observe that the model aggregation and label refinement mutually promote each other during the self-correction process. Concretely, by only employing the MA process, the result shows a gain of 1.74 in terms of mIoU. Meanwhile, the LR process can bring 1.26 improvement. We achieve the best performance by simultaneously introducing these two processes and make an improvement of 2.48 over the baseline result. To better qualitatively understand the effect of our SCHP, Fig. 3.6 shows the visualization of the generated pseudo-masks during the self-correction cycles. Note that all these pseudo-masks are up-sampled to the original size and applied *argmax* operation for better illustration. Label noises like inaccurate boundary, confused fine-grained categories, confused mirror categories, multi-person occlusion are well alleviated or partly resolved during the self-correction cycles. Unsurprisingly, some of the boundaries of our corrected labels are restored to be more smooth than the ground-truth labels. Therefore, the success of our SCHP can attribute to both model aggregation and

Table 3.4: The effect of our proposed model aggregation (MA) and label refinement (LR) strategy is evaluated on LIP `val` set.

Component		Pixel Acc.	Mean Acc.	mIoU
MA	LR			
-	-	87.68	68.79	56.88
✓	-	88.20	71.94	58.62
-	✓	88.14	71.53	58.14
✓	✓	88.42	73.41	59.36

Table 3.5: The SCHP performance on Cityscapes & GTA5 Datasets.

Component	Cityscapes		GTA V	
	mIoU	Δ	*mIoU*	Δ
w/o SCHP	77.82	-	72.73	-
w/ MA	78.75	+0.93	74.09	+1.36
w/ LR	78.01	+0.19	72.84	+0.11
w/ SCHP	78.91	+1.09	74.74	+2.01

label refinement. During the self-correction cycles, the model gets increasingly more robust. Meanwhile, by exploring the dark information from pseudo-masks produced by the enhanced model, the label noises are in turn corrected in an implicit manner.

Influence of Self-Correction Cycles We achieve the goal of self-correction by a cyclically learning scheduler. The number of cycles is thus a virtual hyper-parameter for this process. To make a fair comparison, we maintain the entire training cost unchanged, *i.e.*, the number of total training epochs within self-correction cycles is the same as that of other methods [94, 150]. The performance curves are shown in Fig. 3.8. It is evident that the performance consistently improves during the self-correction process, with the largest improvement after the first cycle and tendency saturates in the end. It should be noted that our SCHP can achieve slightly higher performance when applying more training epochs. Here we train SCHP with 10 self-correction cycles, the performance slightly boosts to 59.58 in terms of mIoU. However, we keep the cycle number as 5 to achieve better a trade-off between computational cost and performance. From the performance curve, we also intelligibly demonstrate the mutual benefit of the model aggregation and the label refinement process.

3.3.5 Discussions

How important is the label noise reduction? To verify the importance of reduction of label noise, we directly train the model with the final refined annotations (one-hot label after argmax) derived from the last self-correction cycle. Compared to model training with the original label, the model achieves 58.13 in terms of mIoU, leading to 1.25 performance boost. This result directly validates the importance of the label noise reduction in our

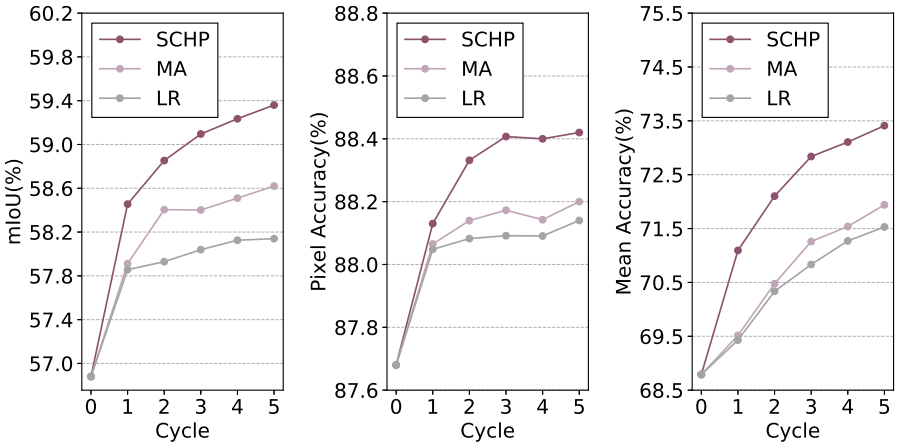


Figure 3.8: Performance curves *w.r.t* different training cycles. The mIoU, pixel accuracy and mean accuracy are depicted in the left, middle and right parts. All experiments are conducted on LIP `val` set.

SCHP.

Can SCHP generalize to clean data? Although SCHP could achieve promising performance improvement by applying the self-correction process to noisy datasets, we reveal that it still works when the ground-truth is relatively clean. To validate our argument, we conduct additional experiments on Cityscapes [79] (high quality annotations) and synthetic dataset GTAV [160] (perfect clean) using Deeplabv3+. We divide the original GTAV dataset into 5 splits, 4 as the training set and 1 as the validation set. As shown in Table 3.5, consistent improvement can be observed. We consider the reasons are as follows. First, the online model aggregation process could serve as a self-generation ensembling, which could lead to better performance and generalization. Second, the online label refinement process benefits from discovering the *dark knowledge* using pseudo-mask instead of the one-hot ground-truth pixel-level label.

How is the effect of SCHP with different noise ratios? To better understanding the effect of SCHP against different noise ratios, we randomly flip total $p\%$ pixels to other classes as artificial label noise on the synthetic dataset, *i.e.*, GTAV, which is with perfect ground-truth labels. In Fig. 3.9, we illustrate the performance corresponding to 0%, 10%, 20% label noise. From the experiments, we observe that the SCHP leads to more performance gains under the scenarios of more label noise. This observation can further prove the effectiveness of our SCHP in coping with noisy data.

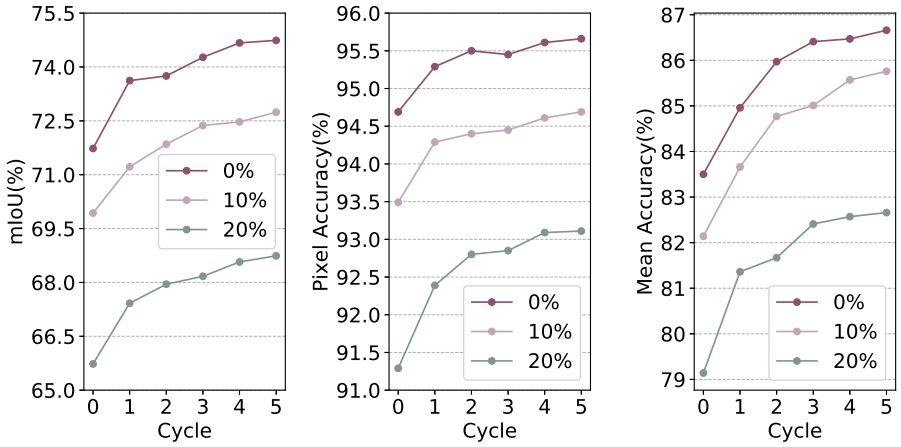


Figure 3.9: Performance *w.r.t* different noise ratios on GTAV dataset.

3.4 Experiments: Multiple-person and Video Human Parsing

Our SCHP is a generic mechanism for tackling human parsing, which can be easily extended for more challenging scenarios. Here we demonstrate the performance of SCHP in addressing multiple-person human parsing tasks and video human parsing tasks. Following the framework described in § 3.2.3, we conduct extensive experiments to evaluate the benefit of SCHP on multiple-person and video human parsing benchmarks.

3.4.1 Experiment Settings

Datasets For the multiple-person human parsing task, we conduct experiments on two popular large-scale benchmarks, *i.e.*, Multi-Human Parsing v2 (MHP) dataset [96] and Crowd Instance-level Human Parsing (CIHP) dataset [141]. MHP dataset contains 25,403 manually annotated multiple-person images with 58 fine-grained semantic labels, dividing into 15,403 images for `train` set, 5,000 images for `val` and 5,000 images for `test` set. CIHP dataset contains 38,280 diverse multiple-person images with 19 fine-grained semantic categories, splitting into 28,280 images for `train` set, 5,000 for `val` set and 5,000 images for `test` set.

For the video human parsing task, we employ the Video Instance-level Parsing (VIP) dataset [100]. VIP is a large-scale video-based multi-person human parsing benchmark with 404 videos in total. The category types of the semantic part labels in VIP are identical to those in CIHP. The dataset is divide into 304 sequences for `train`, 50

Table 3.6: Components analysis on `val` set of CIHP.

Components	$mIoU$	$AP^r_{0.5}$	mAP^r	$AP^p_{0.5}$	mAP^p	$PCP_{0.5}$	$mPCP$
Global Branch Only	62.41	44.34	39.58	48.89	44.55	47.74	40.82
+Local Branch	65.09	50.29	44.74	58.67	49.29	55.38	45.89
+ finetuning	65.47	54.71	48.08	62.76	51.05	58.71	48.04
+ SCHP	**67.47**	**58.94**	**52.00**	**65.59**	**52.74**	**61.28**	**50.12**

Table 3.7: Comparison with state-of-the-arts on VIP `val` set. Our SCHP outperforms the other methods by a large margin. Specially, superior AP^r scores at high IoU thresholds are achieved by our method.

Method	$mIoU$	AP^r_h						AP^r					
		0.5	0.6	0.7	0.8	0.9	mean	0.5	0.6	0.7	0.8	0.9	mean
DFF [163]	35.30	89.90	86.40	74.10	-	-	53.20	20.30	15.00	9.80	-	-	20.30
FGFA [99]	37.50	90.60	88.50	81.00	-	-	57.90	24.00	17.80	12.20	-	-	23.00
ATEN [100]	37.90	**90.80**	**86.70**	81.60	-	-	59.90	25.10	18.90	12.80	-	-	24.10
SCHP	**63.19**	88.98	86.23	**82.31**	**71.21**	**33.90**	**67.57**	**57.77**	**52.93**	**46.31**	**34.14**	**12.40**	**51.41**

sequences for `val` set and 50 sequences for `test` set, respectively.

Evaluation Protocols We choose the mean Intersection over Union (mIoU) for measuring instance-agnostic semantic-level human parsing performance. We adopt average precision based on region (AP^r) [161], average precision based on part (AP^p) [95] and percentage of correctly parsed semantic parts (PCP) [96] for evaluating instance-level human parsing performance. We report the mean value (denote with a prefix m) of AP^r, AP^p and PCP at IoU threshold varying from 0.1 to 0.9 with a step size of 0.1, and the value at 0.5 IoU threshold (denote with subscript 0.5) is also reported. For video human parsing, beyond the above metrics, the mean value of average precision is additionally utilized to measure the whole human instance mask, denoted as AP^r_h [100].

Implementation Details For the human detector, we adopt the implementation of Mask R-CNN [162]. In our experiments, the human detector is further fine-tuned on the corresponding datasets by transforming the human masks into bounding box annotations. During testing, bounding boxes with confidence scores greater than 0.7 are considered as candidates. Then, a spatial area threshold of 0.01 is utilized to filter out some tiny candidate boxes that usually could be considered as background noise. For the local and global human parsing branches, we adopt the same network architecture and configuration as SCHP used in the single-person human parsing task. All input images or video frames are resized to 473×473 before feeding into the branches. In all experiments, we train all three single human parsing branches for 150 epochs with a total batch size of 32.

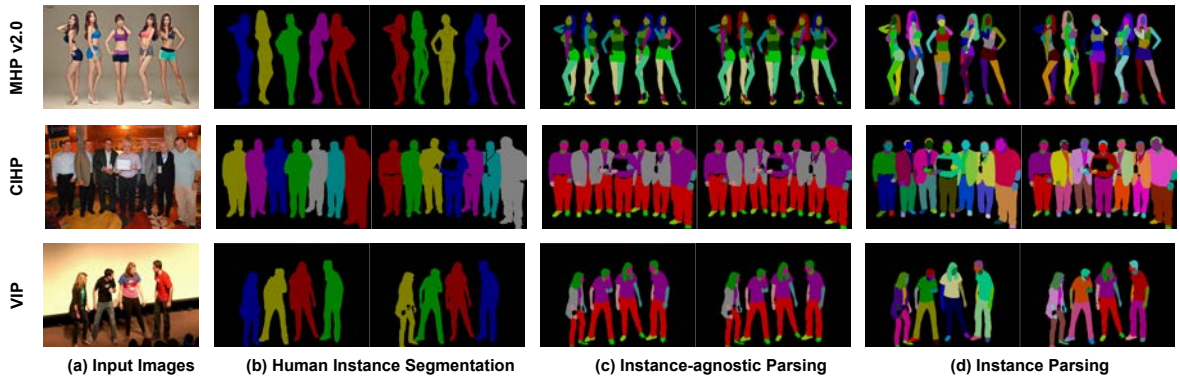


Figure 3.10: Visualization results on MHP v2.0, CIHP and VIP `val` sets. All our results are depicted on the `left` part of each pair, while corresponding ground-truth labels are shown on the `right` side.

3.4.2 Quantitative Results

Multiple-person human parsing In Table 3.9, we report SCHP performance on MHP `val` set comparing with competing state-of-the-art methods. It can be observed that the $mIoU$ of our SCHP significantly outperforms the state-of-the-art approaches by more than 4.1, which can well validate the effectiveness of our SCHP for multiple-person human parsing. Besides, our proposed method also achieved superior results on the instance-aware metric, outperforming NAN-CRF [97] and Parsing R-CNN [164] in terms of mAP^p and $PCP_{0.5}$, respectively. In Table 3.8, following previous practice, we report the performance of $mIoU$, $AP^r_{0.5}$ and mAP^r results on CIHP dataset. Our SCHP reaps quite promising results and outperforms all other methods. Comparing with the former state-of-the-art, our SCHP defeats BraidNet [149] in terms of $mIoU$(66.29 *vs.* 60.62) and mAP^r(51.08 *vs.* 43.59) by a large margin.

To further analyze the effect of each component in our multiple-person human parsing framework, ablation experiments are conducted on CIHP benchmark, as illustrated in Table 3.6. We can observe that when combining both global and local branches, the model gets the best performance comparing to the one with only global or local branches. This could be considered as a kind of intra-model ensembling with the same network architecture but different input sources. The global branch takes the whole image as the input and copes with the instance-agnostic parsing under the complex scenes. Meanwhile, the local branches take the cropped images from the human detector as input and thus distinguish the human instances from each other. By fine-tuning the human detector, a great performance boost from instance-aware metrics, *i.e.*, AP^r, AP^p and PCP can be

Table 3.8: Comparison with state-of-the-arts on CIHP dataset.

Method	subset	$mIoU$	$AP^r_{0.5}$	mAP^r
Parsing R-CNN [164]	val	61.10	-	-
PGN [141]	test	55.80	35.80	33.60
Graphonomy [98]	test	58.58	-	-
CE2P [94]	test	59.50	48.69	42.83
BraidNet [149]	test	60.62	48.99	43.59
SCHP	val	**67.47**	**58.94**	**52.00**
SCHP	test	**66.29**	**57.58**	**51.08**

Table 3.9: Comparison with state-of-the-arts on MHP val set.

Method	$mIoU$	$AP^p_{0.5}$	mAP^p	$PCP_{0.5}$	$mPCP$
Mask R-CNN [6]	-	14.50	33.51	25.12	-
NAN [96]	-	24.83	42.77	34.37	-
NAN-CRF [97]	-	27.92	44.95	36.63	-
Parsing R-CNN [164]	41.80	32.50	42.70	47.90	-
Graphonomy [98]	34.05	-	-	-	-
CE2P [94]	41.11	34.47	42.70	43.77	41.06
SCHP	**45.21**	**35.10**	**45.25**	**48.02**	**42.30**

observed. Most importantly, with the benefit of SCHP, great improvements are achieved.

Video human parsing To further validate the generalization ability of our solution, we extend the framework from image multiple human parsing to video human parsing tasks. In Table 3.7, we report the results on VIP val set comparing with other state-of-the-art methods. Our proposed SCHP maintains a large-margin leading edge on all main metrics. Comparing with state-of-the-art video human parsing method ATEN [100], superior improvements of +25.29, +7.67 and +30.31 in terms of $mIoU$, mAP^r_h and mAP^r are achieved, respectively. Specially, to better demonstrate the superiority of our proposed method, we also report the results on AP^r_h and AP^r at high IoU thresholds 0.8 and 0.9. It can be observed that our proposed method gets 12.40 AP^r at IoU threshold of 0.9. This reflects a promising performance and higher parsing accuracy of ours, considering that FGFA [99] and ATEN [100] only get 12.20 and 12.80 AP^r at IoU threshold of 0.7.

3.4.3 Qualitative Results

Several challenging images or frames from MHP, CIHP and VIP datasets are depicted in Fig. 3.10. Additionally, we visualize the results and corresponding annotations on the same row, including the human instance masks, instance-agnostic parsing results and instance parsing results. In the first row of Fig. 3.10 (c), noisy annotations are provided, *i.e.*, four of five models' hair (colored in blue) are incorrectly annotated as *cap/hat* category (colored in red). Nevertheless, based on the high robustness of our proposed method, parsing results with higher quality can be provided. Not only the

semantic categories of each pixel can be identified correctly, but more smoothly edges are maintained. Similar results can be observed from many other images on both multiple human parsing and video human parsing datasets. This reveals our proposed method has the superiority of self-correction and training with some noisy data. Results from Fig. 3.10 also demonstrate the effectiveness of coping with many challenging conditions, such as crowed senses, appearance variability, occlusions and *etc*.

3.5 Summary

In this section, we tackle the problem of human parsing task under learning with label noise scenario, which is never explored before. From a new perspective, we unravel the problem by exploring the effect of model aggregation, label refinement and their interaction. Based on our investigation, we present a simple yet effective, generic, model-agnostic mechanism called SCHP for human parsing task to deal with the label noise during training process by self-correction. We validate the effectiveness and generalization ability of our method for all human parsing tasks, including single-person human parsing, multi-person human parsing and video human parsing. Benefiting from our SCHP, a consistent performance boost is observed for all three tasks, leading to the new state-of-the-art performance for all large-scale human parsing benchmarks. Incorporating advanced techniques, our overall system ranks 1st for all human parsing tracks for the LIP challenge at CVPR2019. We hope our work could serve as a start point and facilitate future research.

CHAPTER 4

META PARSING NETWORKS: TOWARDS GENERALIZED FEW-SHOT SCENE PARSING WITH ADAPTIVE METRIC LEARNING

4.1 Preface

Semantic segmentation [8, 137] aims at assigning a unique semantic label to each pixel in the given image. Recently, deep learning [5, 156] has significantly advanced the development of the semantic segmentation. Many promising architectures based on fully convolutional networks are proposed to tackle such a challenging task, including FCN [137], PSPNet [7] and DeepLab series [8, 71]. However, the success of advanced architectures heavily relies on a large number of training images with pixel-level annotations, which are often expensive to be obtained. Moreover, the current semantic segmentation scheme is usually close-set based, *i.e.*, all categories are pre-defined before training, leading to the learned segmentation models cannot be generalized to the novel (unseen) categories.

To tackle the aforementioned issues and enable the segmentation models to equip with good generalization ability to unseen categories, meta-learning provides promising solutions. In general, meta-learning, also known as learning to learn, targets on learning new concepts or skills fast with only a few training samples. Some few-shot segmentation approaches [66, 101, 106] have been proposed to expand semantic segmentation with

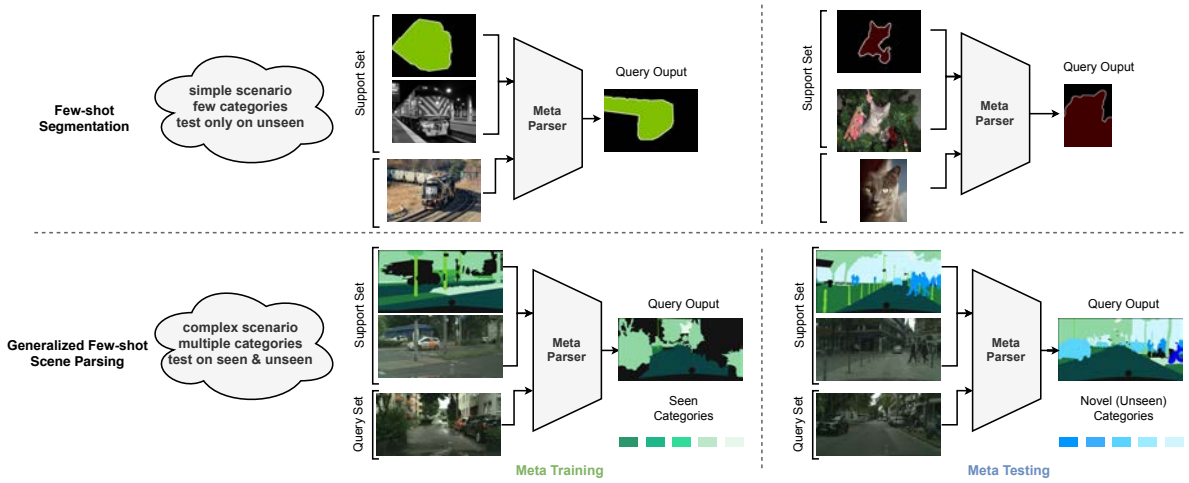


Figure 4.1: Compared to conventional few-shot segmentation task, the generalized few-shot scene parsing aims to segment complex scene scenario with multiple visual categories, where both seen and unseen categories are simultaneously considered. During meta-training stage, we first train the meta parser with the annotated images only on *seen* categories (e.g., road and vegetation), as indicated by the green colors. During meta-testing stage, given only one annotation image (one-shot) as guidance, the learned meta parser is then applied to segment both *seen* categories and *unseen* ones (e.g., car and person , as indicated by the blue colors.). Our target is to learn a meta parser that can generalize to both *seen* and *unseen* categories.

the ability to segment unseen categories with few annotation samples. A standard setting of the current few-shot segmentation usually follows: feed one support image and one query image into the network simultaneously and segment the query image using the support image as the guidance. Although impressive progress has been made, the current explorations are only solving segmentation with simple objects (usually one or two categories) instead of scene parsing with a complex fully annotated image, which is actually a more realistic yet challenging task.

In this work, we advance the few-shot segmentation paradigm towards a more challenging yet general scenario, *i.e.*, generalized few-shot scene parsing. Under such a setting, we take a fully annotated image as guidance and perform the segmentation to all pixels in a query image (as illustrated in Fig. 4.1). Our mission is to study a robust segmentation network from the meta-learning perspective so that pixels from both seen and unseen categories could be well recognized. Compared with object-based few-shot segmentation, our setting raises two additional challenges. First, we need to explore more effective metric learning solutions to generate discriminative prototypes for dozens of categories in the given support image, rather than only one or two categories considered

by the object segmentation setting. Second, the segmentation model is expected to produce class-aware feature representations, making the model have a strong generalization ability to multiple unseen categories, simultaneously.

To tackle these problems, in this work, we present a generic framework, named Meta Parsing Networks (MPNet). Our MPNet mainly contains two simple yet effective modules, *i.e.*, Adaptive Deep Metric Learning (ADML) module and the Contrastive Inter-class Distraction (CID) module. To be specific, the ADML module is responsible for modeling the relationship between the pixels in the query image and the annotated ones in the support image. We adaptively leverage non-local operations to generate discriminative prototypes of each category within the support image, which is then employed to learn a deep metric comparison with the query image. Moreover, we introduce a CID module, which can be considered as a regularization item to encourage the feature discrepancy of different categories to be as larger as possible. In this way, our MPNet will be imposed to produce class-sensitive feature representations, resulting in better generalization ability to both seen and unseen categories, accordingly.

Our MPNet is a generic framework for performing the generalized few-shot scene parsing task. We conduct experiments on two newly constructed generalized few-shot scene parsing benchmarks, called *GFSP-Cityscapes* and *GFSP-Pascal-Context*. Extensive ablation studies and comparisons well demonstrate the effectiveness and generalization ability of our proposed MPNet. Even though this work takes one step closer to the generalized few-shot scene parsing task, there is actually still a long way to go. We hope our efforts can motivate more researchers to design more robust meta-learning-based algorithms and benefit the research of few-shot scene parsing in the future.

4.2 Task Definition

Few-shot learning splits the input data into an annotated support set, which provides supervised signal to guide the learning process, and an unannotated query set on which to do the task. Former works [10, 14] tend to address the few-shot learning problem by re-casting it into a meta-learning paradigm. The meta-learning paradigm includes *meta-train* and *meta-test* as two phases. Meanwhile, the class sets are disjoint between $\mathscr{C}_{seen}$ and $\mathscr{C}_{unseen}$. In the *meta-train* phase, to simulate and encourage a fast adaptation and task generalization ability, the *episodic training* scheme [10] is adopted. For each episode, the model is trained only using a sampled subset of $\mathscr{C}_{seen}$. In the *meta-test* phase, the generalization ability of the learned model is examined on samples from $\mathscr{C}_{unseen}$.

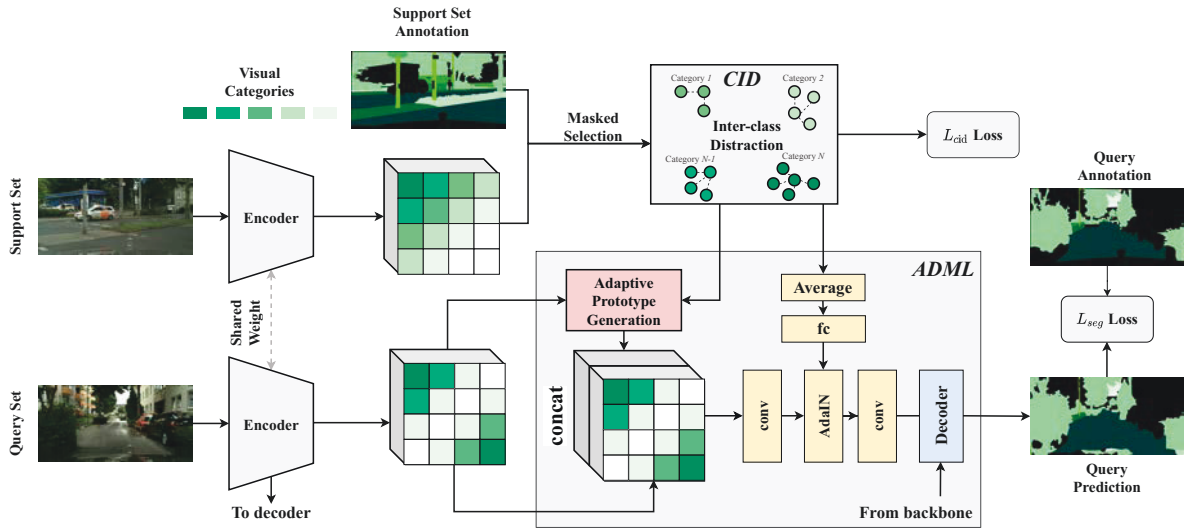


Figure 4.2: An overview of our MPNet. The ADML module aims to adaptively *learn* a transferable deep metric for dense comparison between support and query images. The CID module aims to encourage the feature discrepancy of different categories.

Inspired by the conventional setting of few-shot learning, we introduce and formalize the protocols of generalized few-shot scene parsing as illustrated in Fig. 4.1. Similar to few-shot object segmentation [66, 165], the training set $\mathscr{D}_{train}$ is only constructed from $\mathscr{C}_{seen}$. Differently, the testing set $\mathscr{D}_{test}$ is constructed from both $\mathscr{C}_{seen}$ and $\mathscr{C}_{unseen}$ rather than $\mathscr{C}_{unseen}$ as adopted by few-shot object segmentation. We aim to leverage the $\mathscr{D}_{train}$ to learn a meta parser, which can be well generalized to perform open-set evaluation. We follow the common practice of few-shot learning, and conduct episodic training/testing on $\mathscr{D}_{train}$ and $\mathscr{D}_{test}$. Specifically, every episode is constructed by a set of support images $\mathscr{S}$ and a set of query images $\mathscr{Q}$. In general, the support set $\mathscr{S}$ has 1 to K $\langle image, mask \rangle$ pairs, which provides guidance to segment the unlabeled image in $\mathscr{Q}$. Different from the few-shot object segmentation that each support image represents one unique semantic class, several classes often appear within one support image in our setting. Additionally, dozens of classes may span on multiple support images. Therefore, given the $\mathscr{D}_{train} = \{(\mathscr{S}_i, \mathscr{Q}_i)\}^{N_{train}}$, our mission is to apply the learned knowledge from $\mathscr{D}_{train}$ to perform scene parsing on $\mathscr{D}_{test} = \{(\mathscr{S}_i, \mathscr{Q}_i)\}^{N_{test}}$, where N_{train} and N_{test} are the number training/testing episodes, respectively. Based on the number of pairs given in the support set, the few-shot open-set scene parsing can be instantiated as K-shot segmentation learning task, accordingly.

4.3 Meta Parsing Networks

Overview We propose the Meta Parsing Networks (MPNet) to tackle the generalized few-shot scene parsing in this work, as shown in Fig. 4.2. The key idea of our MPNet is to parse the scene by exploring the dense feature comparison from the metric learning perspective *without the need for performing further fine-tuning steps*. Without loss of generality, we begin with the illustration of our model in a one-shot setting. Our MPNet consists of two basic modules: the Adaptive Deep Metric Learning module (ADML) and the Contrastive Inter-class Distraction (CID) module. The ADML module aims to *learn* a transferable deep metric for dense comparison between support and query images. Instead of comparing the feature based on a human pre-defined metric (*e.g.*, as cosine or euclidean), we make the network adaptively learn a generalizable deep metric which adapts fast to novel unseen class. In addition, to better impose the feature discrepancy of different categories to be as larger as possible, we further introduce the CID module to conduct the contrastive learning inspired by Maximum Mean Discrepancy [166]. Moreover, our MPNet can be easily extended from one-shot learning to K-shot learning without much computation overhead.

4.3.1 Adaptive Deep Metric Learning

Support and Query Embedding In one-shot setting, one support image and its corresponding mask are leveraged to guide the segmentation of the query image in each training episode. Both the support and the query images are first encoded into feature embeddings through a dedicated deep encoder as shown in Fig. 4.2. The deep encoder aims to harvest comprehensive representations from convolutional neural networks for further deep metric learning. Denote the output 2D embedding maps for the support image and the query image are $\boldsymbol{F}^{\mathscr{S}} \in \mathbb{R}^{H\times W\times C}$ and $\boldsymbol{F}^{\mathscr{Q}} \in \mathbb{R}^{H\times W\times C}$, where H, W and C are the height, the width and the channel numbers, respectively. In our implementation, we take the DeeplabV3 [8] as the basic segmentation framework of our MPNet and the ResNet101 [5] is employed as the backbone.

Adaptive Prototype Generation (APG) As a full image, it usually contains complex categories from both *stuff*(*e.g.*, sky, road, building, etc.) and *things*(*e.g.*, person, car, bus, etc.) with different appearances from scales, shapes or locations. We need to firstly acquire the class-specific feature embeddings from the support image for performing the following deep metric comparison. Previous practices [66, 165] adopt class-aware average pooling to obtain the representative class-specific embedding as the prototype of

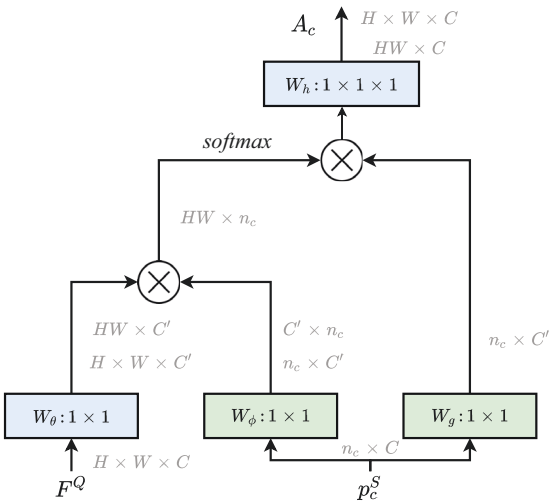


Figure 4.3: Illustration of the Adaptive Prototype Generation.

a class. However, we argue that simply pooling all the pixel-wise features into one single vector for a specific category will eliminate the rich diversity information among different pixels, which is harmful for those categories with similar appearances. This situation will get even worse in the scene parsing scenario where all the pixels from multiple categories require to be joint considered. To fully exploit the rich diversity information embedded in the class-specific pixels, we propose the APG to measure the similarities between the embedding maps $\boldsymbol{F}^{\mathcal{S}}$ and $\boldsymbol{F}^{\mathcal{Q}}$, using pixel-to-pixel comparison. Our APG is motivated from the non-local operations adopted by current advanced self-attention techniques [167, 168]. Concretely, given a support image embedding maps $\boldsymbol{F}^{\mathcal{S}}$ and its ground-truth mask $M^{\mathcal{S}}$. The corresponding embedding collection of category c is selected via

$$\boldsymbol{p}_c^{\mathcal{S}} = \{\boldsymbol{F}_{(x,y)}^{\mathcal{S}} | \mathbb{1}[M_{(x,y)}^{\mathcal{S}} = c]\}, \tag{4.1}$$

where (x, y) indicates the location and $\mathbb{1}(\cdot)$ is an indicator function to select its corresponding features. The selected features are $\boldsymbol{p}_c^{\mathcal{S}} \in \mathbb{R}^{n_c \times C}$ where $n_c = \sum_{(x,y)} \mathbb{1}[M_{(x,y)}^{\mathcal{S}} = c]$ indicates the number of pixels. During the adaptive feature prototype generation process as shown in Fig. 4.3, to save computation cost and memory usage, we first project the selected support embeddings $\boldsymbol{p}_c^{\mathcal{S}}$ into key maps $\boldsymbol{k}^{\mathcal{S}}$ and value maps $\boldsymbol{v}^{\mathcal{S}}$ by 1×1 convolutions W_ϕ, W_g. Also the query features $\boldsymbol{F}^{\mathcal{Q}}$ are also mapped into the key maps $\boldsymbol{k}^{\mathcal{Q}}$ by 1×1 convolutions W_θ

$$\boldsymbol{k}^{\mathcal{S}} = W_\phi(\boldsymbol{p}_c^{\mathcal{S}}), \quad \boldsymbol{v}^{\mathcal{S}} = W_g(\boldsymbol{p}_c^{\mathcal{S}}), \quad \boldsymbol{k}^{\mathcal{Q}} = W_\theta(\boldsymbol{F}^{\mathcal{Q}}), \tag{4.2}$$

where the channel number is all reduced from C to the C'. Then, adaptive soft attentions are first computed by the matrix multiplication between all query key in $\boldsymbol{k}^{\mathcal{Q}}$ and the selected embeddings key $\boldsymbol{k}^{\mathcal{S}}$. Finally, the adaptive prototype for class c is retrieved by applying the softmax normalized weighted summation

$$\boldsymbol{v}_i^{\mathcal{Q}} = \frac{1}{Z}\sum_{j=1}^{n_c} f\left(\boldsymbol{k}_i^{\mathcal{Q}}, \boldsymbol{k}_j^{\mathcal{S}}\right)\boldsymbol{v}_j^{\mathcal{S}}, \tag{4.3}$$

where i, j are the location index of embedding in query maps and the selected support embedding collection, respectively. The normalizing factor is defined as $Z = \sum_{j=1}^{n_c} f\left(\boldsymbol{k}_i^{\mathcal{Q}}, \boldsymbol{k}_j^{\mathcal{S}}\right)$. The similarity calculation is implemented by dot product as

$$f\left(\boldsymbol{k}_i^{\mathcal{Q}}, \boldsymbol{k}_j^{\mathcal{S}}\right) = \frac{1}{\sqrt{C'}}\exp\left(\boldsymbol{k}_i^{\mathcal{Q}} \cdot \boldsymbol{k}_j^{\mathcal{S}}\right). \tag{4.4}$$

The query value $\boldsymbol{v}^{\mathcal{Q}}$ is further mapped back to the original channel number by 1×1 convolution W_h, obtaining the adaptive generated prototype embedding $\boldsymbol{A}_c$, accordingly.

Adaptive Feature Comparison After the adaptive prototype generation, we concatenate the generated prototype embedding maps $\boldsymbol{A}_c$ with the query embedding maps $\boldsymbol{F}^{\mathcal{Q}}$. Then, the concatenated feature maps go through several convolutional blocks to learn a deep metric for comparing whether the query features belong to the specific category or not. Such a feature comparison mechanism makes the meta-parser equipped with the function of distinguishing the pixels related to the selected category from all pixels in the query image. To further enhance the comparison process and make the produced features adapted to class-specific parsing, we additionally introduce the Adaptive Instance Normalization (AdaIN) layer motivated by the recent progress in style transfer [135], as shown in the ADML module in Fig. 4.2. To be specific, we parameterize the deep metric comparison by the guidance from the mean category prototype $\overline{\boldsymbol{p}}_c^{\mathcal{S}}$. We forward the $\overline{\boldsymbol{p}}_c^{\mathcal{S}}$ by fully connected layers to learn a set of affine parameters $\boldsymbol{\gamma}, \boldsymbol{\beta}$ in the instance normalization layer, where $\boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^C$. In this way, the meta-parser will be encouraged to learn class-sensitive feature embeddings for conducting the comparison, leading to a better parsing result for the selected category.

Different from the few-shot object segmentation that only considers one or two categories, the scene parsing task requires the meta-parser to process dozens of categories, simultaneously. To efficiently and systematically deal with a variable number of categories, we dynamically instantiate once for each category with our adaptive deep metric learning module as shown in Fig. 4.4. Concretely, the inputs to the ADML module for each category are i) the query image embedding maps $\boldsymbol{F}^{\mathcal{Q}}$ and ii) the selected embedding

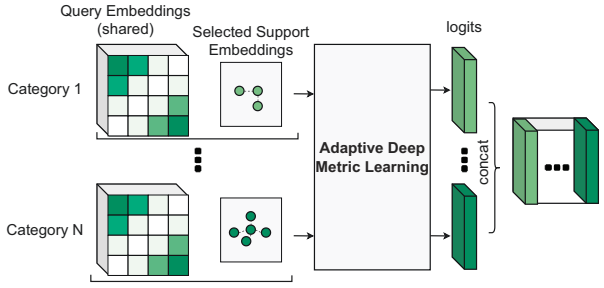


Figure 4.4: Dynamic instantiation for multiple categories .

collection $\boldsymbol{p}_c^{\mathcal{S}}$. We cyclically conduct the adaptive feature comparison for each category until all the categories in the given support image are considered. The feature maps calculated by each comparison are concatenated together to conduct the final segmentation prediction for all categories. Although the ADML module needs to run multiple times, the major computation cost actually occurs in extracting the features from the backbone which allows our MPNet scale well to complex scene scenarios. Formally, a one-dimensional feature map of logits is extracted for each category. We stack them together as $\tilde{\boldsymbol{M}}^{\mathcal{Q}} \in \mathbb{R}^{H\times W\times C}$, feeding to the classifier and apply the cross-entropy loss for the optimization

$$\mathscr{L}_{\text{seg}} = -\frac{1}{HW}\sum_{x,y}\sum_{c}\mathbb{1}\left[M^{\mathcal{Q}}_{(x,y)}=c\right]\log\tilde{\boldsymbol{M}}^{\mathcal{Q}}_{(x,y),c}. \tag{4.5}$$

4.3.2 Contrastive Inter-class Distraction

Different from the few-shot object segmentation, there exist dozens of complex visual categories with different shapes and location in the support image. The feature ambiguity for those categories with similar appearances will bring additional negative effects to the adaptive deep metric learning. Intuitively, if the feature discrepancy of different categories from the support image could be enlarged, the subsequent adaptive deep metric learning will be implicitly benefited. To this end, we further propose a Contrastive Inter-class Distraction (CID) module to augment the ADML for leaning a better meta parser. Our CID is partly inspired by the Maximum Mean Discrepancy [166], which models the difference between two distributions in the Reproducing Hilbert Kernel Space. Without the need of knowing the concrete semantic meaning of two visual categories c_m and c_n, we select their corresponding feature collections from the support feature embeddings by Eq (4.1), i.e., $\boldsymbol{p}^S_{c_m}$ and $\boldsymbol{p}^S_{c_n}$. Then we calculate the difference between

these two feature distribution by

$$\begin{aligned}\mathscr{L}^{mmd}_{(c_m,c_n)} = &\frac{1}{n^2_{c_m}}\sum_{i=1}^{n_{c_m}}\sum_{j=1}^{n_{c_m}} k\left(\boldsymbol{p}^S_{c_m,i},\boldsymbol{p}^S_{c_m,j}\right) + \frac{1}{n^2_{c_n}}\sum_{i=1}^{n_{c_n}}\sum_{j=1}^{n_{c_n}} k\left(\boldsymbol{p}^S_{c_n,i},\boldsymbol{p}^S_{c_n,j}\right) \\ &- \frac{2}{n_{c_m}n_{c_n}}\sum_{i=1}^{n_{c_m}}\sum_{j=1}^{n_{c_n}} k\left(\boldsymbol{p}^S_{c_m,i},\boldsymbol{p}^S_{c_n,j}\right),\end{aligned} \tag{4.6}$$

where we choose the radial basis function kernel with bandwidth σ as the kernel function k, defined as

$$k\left(\mathbf{x},\mathbf{x}'\right) = \exp\left(-\frac{\left\|\mathbf{x}-\mathbf{x}'\right\|_2^2}{2\sigma^2}\right). \tag{4.7}$$

Suppose there totally exists N classes in the support image set, we calculate the distribution difference pair-wisely and sum all these item together to get our CID loss

$$\mathscr{L}_{cid} = -\frac{1}{N(N-1)}\sum_{c=1}^{N}\sum_{\substack{c'=1\\c\neq c}}^{N}\mathscr{L}^{mmd}_{(c,c')}. \tag{4.8}$$

By optimizing the CID loss, we explicitly model the inter-class separability. Our CID loss is non-parametric and has no assumption in knowing the semantic meaning of each class. These characteristics make the CID be general in tackling the generalized few-shot scene parsing task.

4.3.3 Meta-training & Meta-testing

Meta-training Our MPNet is joint trained in an end-to-end manner by considering two loss functions for the optimization, *i.e.*,

$$\mathscr{L}(\boldsymbol{\theta}) = \mathscr{L}_{seg} + \lambda\mathscr{L}_{cid}, \tag{4.9}$$

where $\mathscr{L}_{seg}$ and $\mathscr{L}_{cid}$ are the loss functions for performing the adaptive deep metric learning (as defined in Eq (4.5)) and contrastive inter-class distraction (as defined in Eq (4.8)), respectively. λ is the weight of the discrepancy regularization term, which is experimentally set as 0.1 in this work. $\boldsymbol{\theta}$ indicates the learned parameters of our MPNet. As shown in the left of Fig. 4.1, only classes from the *seen* categories (as indicated by the green colors) are selected during the training and other classes (as indicated by the black color) that play as the *unseen* categories are masked, thus do not participate in optimization. For each training episode, we construct pairs of images as the inputs, one of which serves as the support image and the other one is the query image accordingly.

Meta-testing As shown in the right of Fig. 4.1, both the *seen* and the *unseen* (as indicated by the blue colors) categories are considered during the meta-testing stage. For each query image from the testing set, we choose the image that includes the same categories as its support counterpart. Based on the number of support images, we report the results of both 1-shot and 5-shot settings. In addition, to verify the generalization ability of our MPNet, we also report the results on both *seen* and *unseen* categories.

4.4 Experiment

In this section, we conduct extensive experiments to evaluate our Meta Parsing Networks, trying to answer one central question – *how is the generalization ability of the Meta Parsing Networks?*

4.4.1 Generalized Few-shot Scene Parsing Benchmarks

As currently there are no datasets targeting at the Generalized Few-shot Scene Parsing (GFSP) task, we propose two benchmarks, namely *GFSP-Cityscapes* and *GFSP-Pascal-Context* to evaluate the performance of MPNet. The two benchmarks are built upon the Cityscapes Dataset [79] which focuses on the urban scene parsing scenario and Pascal-Context Dataset [82] which focuses on the realistic scenes.

Table 4.1: *GFSP-Cityscpaes* benchmark splits. We only list the unseen categories, all rest are seen categories.

GFSP-Cityscapes	Unseen Categories
split1	$road, wall, traffic\ sign, sky, rider, bus, motorcycle$
split2	$sidewalk, fence, pole, terrain, car, bicycle$
split3	$building, traffic\ light, vegetation, person, truck, train$

GFSP-Cityscapes Benchmark The Cityscapes dataset consists of images covering 19 urban scene categories. To simulate an generalized few-shot scene parsing setup, we cyclically divide all categories into three different $\{seen/unseen\}$ split as listed in Table 4.1. During the meta-training, both the support and query images are sampled from the training set. We mask all the unseen categories in the original annotation thus only the seen categories are considered for training. During the meta-testing, we take images from the original validation set as queries and both seen and unseen are required to be predicted. Previous practice [66] demonstrates that it easily leads to relative large variance by randomly sampling only few episodes for the meta-testing. To alleviate the

unstable evaluation as much as possible, in our experiments, we report 5 runs using 5 different fixed support images sets. Note that the support images during meta-testing are never seen during meta-training.

GFSP-Pascal-Context Benchmark The Pascal-Context dataset is a challenging scene parsing dataset that contains 59 semantic classes. The training set and validation set consist of 4,998 and 5,105 images, respectively. Upon this dataset, we set the co-existing 14 categories in cityscapes dataset as the unseen categories, the rest classes as the seen categories. By this seen/unseen split, we can evaluate the our generalized few-shot scene parsing performance not only on the GFSP-Pascal-Context Benchmark itself, but also give possibility for evaluating the cross-benchmark transfer setup from GFSP-Pascal-Context to GFSP-Cityscapes.

Evaluation Metric Some former few-shot object segmentation practices [169, 170], they ignore the image categories and only calculate the class-agnostic mean of foreground IoU and background IoU over all test images. However, in the generalized few-shot scene parsing task, there may have different classes in the support image, the number of pixels in different classes is not balanced. Ignoring the class categories will lead to a biased performance for class with more pixels. Thus, in all of our experiments, we choose the mean intersection-over-union (mIoU) metric over seen/unseen categories to better evaluate the performance. Results are reported by averaging over 5 runs.

Extension from one-shot to few-shot To maintain simplicity, we describe our MPNet only for one-shot setting. However, one can easily extend MPNet from one-shot to K-shot setting. Comparing with one-shot learning which has only one support image, K-shot learning contains K images in the support set which contains more abundant guidance information. For the K-shot testing, previous method [101] applies one-shot method independently to each support example and fuse individual predicted results at the image level. In contrast, we can effectively fuse the information from multiple support examples with only once forward thanks to our adaptive prototype generation mechanism. Particularly, we merge the selective embedding collection $\boldsymbol{p}_c^{\mathbf{S}}$ in Eq 4.1 together to enlarge the feature samples.

Baseline Methods. Since the generalized few-shot scene parsing is a new task, we construct the following baselines to examine the effectiveness of our proposed MPNet.

- *Mask Siamese.* Siamese Network [11] predicts whether two inputs belong to the same class or not, showing good performance on few-shot image classification. Followed by [101], we adapt this method using the extracted dense features to train the network for pairwisely dense pixel verification.

Table 4.2: The comparison on *GFSP-Cityscapes* benchmark.

	Method	1-shot				5-shot				Δ
		split1	split2	split3	*mean*	split1	split2	split3	mean	
Seen	Mask Siamese [11]	51.6	52.3	49.7	51.2	51.3	53.0	50.1	51.5	0.3
	Mask Prototype [13]	57.3	56.3	50.1	54.6	59.0	57.5	51.5	56.0	1.4
	MPNet Init	29.8	28.4	27.2	28.5	32.6	31.8	30.9	31.8	**3.3**
	PANet [66]	59.1	**57.8**	51.1	56.0	60.2	59.1	52.6	57.3	1.3
	MPNet(Ours)	**61.8**	57.6	**53.5**	**57.6**	**65.0**	**60.4**	**55.1**	**60.2**	2.6
UnSeen	Mask Siamese [11]	18.4	18.1	20.3	18.9	19.7	18.0	20.9	19.5	0.6
	Mask Prototype [13]	24.3	19.8	21.7	21.9	27.8	23.0	25.1	25.3	3.4
	MPNet Init	22.9	19.8	20.3	21.0	27.1	23.3	24.2	24.9	3.9
	PANet [66]	25.9	20.5	21.0	22.5	28.1	23.5	24.2	25.3	2.8
	MPNet(Ours)	**28.7**	**20.7**	**25.7**	**25.0**	**32.8**	**27.7**	**29.2**	**29.9**	**4.9**

- *Mask Prototype.* Prototypical Networks [13] computes the mean embedding of support images for each class. Followed by [66, 165], we adapt this method to few-shot segmentation by masked average pooling the corresponding class features as the prototype. The query mask is calculated by a cosine metric from each pixel in the query image to every prototype in the support set.

- *MPNet-Init.* In our MPNet implementation, we adopt the pre-trained model on ImageNet [156]. Although all images in our meta-testing phase are never seen during the meta-training phase, some unseen categories may have partial relation or overlap with some ImageNet visual concepts. To better disentangle this effect, we fixed the pre-trained model and only the parameters from the ADML module are updated during the meta-training phase.

- *PANet.* PANet [66] is one of the state-of-the-art approaches in few-shot object segmentation with publicly available code. We directly extend this method for our generalized few-shot scene parsing setting. Compared with our MPNet, this work adopts a fixed cosine metric and the relation among all categories are not explored. Note that all methods adopt the same encoder as MPNet to make a fair comparison.

4.4.2 Generalization Ability of MPNet

We conduct extensive experiments to make a comparison between our MPNet and other baseline methods. For all the referred Tables in this subsection, the performance is reported as the mean mIoU with five runs and Δ denotes the performance gain between 1-shot and 5-shot learning.

Table 4.3: The comparison on *GFSP-Pascal-Context* benchmark.

Method	Seen Categories			Unseen Categories		
	1-shot	5-shot	Δ	1-shot	5-shot	Δ
Mask Siamese [11]	24.4	24.6	0.2	17.1	17.4	0.3
Mask Prototype [13]	30.8	32.6	1.8	24.2	26.4	2.2
MPNet Init	28.1	30.1	2.0	23.4	25.7	2.3
PANet [66]	32.1	33.8	1.7	25.2	27.7	2.5
MPNet (ours)	**35.3**	**37.9**	**2.6**	**28.7**	**32.2**	**3.5**

Table 4.4: The comparison of cross-domain experiments from *GFSP-Pascal-Context* to *GFSP-Cityscapes* benchmark.

Method	Unseen Categories		
	1-shot	5-shot	Δ
Mask Siamese [11]	13.1	13.4	0.3
Mask Prototype [13]	17.9	20.7	2.8
MPNet Init	16.4	19.3	2.9
PANet [66]	18.7	21.3	2.6
MPNet (ours)	**20.3**	**23.6**	**3.3**

Does the MPNet generalize on unseen categories? We show the qualitative results of our MPNet both under 1-shot and 5-shot settings on *GFSP-Cityscapes* benchmark in Table 4.2. As can be observed, MPNet consistently outperforms the other models by a large margin across different seen/unseen splits, especially on the unseen categories. Concretely, for one-shot setting, MPNet outperforms the previous state-of-the-art model PANet by 1.6% on seen categories and 2.5% on unseen categories, respectively. Our MPNet achieves better generalization ability by exploring, 1) an adaptively learned deep metric for dense pixel comparison and 2) discovering the inter-class relation simultaneously. In comparison, we argue that a learned deep metric generalizes better than a fixed metric. This finding is also consistent with [14]. By endowing a learned metric, MPNet outperforms the Mask Siamese with L1 metric (57.6 *vs.* 51.2), Mask Prototype (57.6 *vs.* 54.6) and PANet (57.6 *vs.* 56.0) with cosine metric for dense pixel comparison. Besides, other methods only target on object segmentation by applying class-agnostic segmentation separately. MPNet overcomes this weakness by discovering the inter-class relationship simultaneously. Table 4.3 shows the comparison on the *GFSP-Pascal-Context* benchmark. We can observe that MPNet still outperforms the other baseline models and PANet (3.2% on seen categories and 3.5% on unseen categories). Note that, on both benchmarks the performance gain on unseen categories is higher than that on seen categories, which reveals the superiority on discovering novel classes.

Does multiple support images benefit? In the real-world scenario, users may provide multiple annotated images (K-shot) as a more abundant query set. As shown in Table 4.2, the performance gain between 1-shot and 5-shot is larger than other methods (4.9 *vs.* 2.8 on unseen categories). This comparison well demonstrates the advantage of our MPNet in gathering useful guidance information when more support information is available. The trend is also consistent as shown in Table 4.3. We further investigate the performance of MPNet on *GFSP-Cityscapes* benchmark with more support images as

illustrated in Fig. 4.5. It is evident that the performance consistently improves with more support images, with the large gain at the beginning while tendency saturates at the end. Note that all images in the support set are only used as the guidance information rather than fine-tuning on these images.

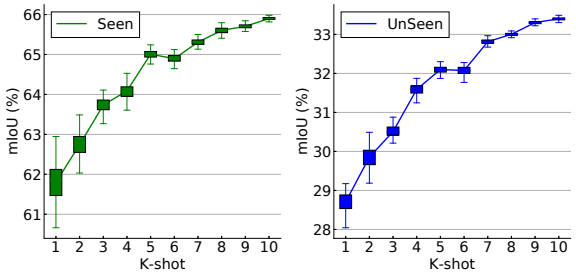


Figure 4.5: Performance *w.r.t* the number of support images (K-shot).

Generic Embedding *vs.* Learned Embedding We evaluate two kinds of embedding for our adaptive deep metric learning, *i.e.,* 1) we freeze the ImageNet pre-trained model during training as the generic embedding. 2) end-to-end learning the embedding on our benchmarks. As shown in Table 4.2, the learned embedding outperforms the generic embedding (57.6 *vs.* 28.5 on seen categories and 25.0 *vs.* 21.0 on unseen categories). Thus we conclude that although the MPNet parses the unseen categories by a deep-metric based comparison, the end-to-end learned embedding is still necessary.

Does the MPNet generalize across domain? Further, distinct from the previous experiments that focus on evaluating in-domain model generalization, we estimate the MPNet to reveal the cross-domain generalization ability in Table 4.4. The model is trained on seen categories on *GFSP-Pascal-Context* and tested on unseen categories on *GFSP-Cityscapes*. Still the MPNet achieves better performance compared with other methods, which verifies the generalization ability of the proposed MPNet.

4.4.3 Further Analysis of MPNet

Here we conduct comprehensive ablation analysis in Table 4.5 to uncover the effectiveness of the proposed modules of MPNet. All ablations are based on 1/5-shot generalized few-shot scene parsing performances on *GFSP-Cityscapes* in the first seen/unseen split setup.

Adaptive Prototype Generation Instead of merely performing class-aware average pooling, the APG module is proposed to better maintain the rich diversity information

Figure 4.6: Visualization results of MPNet. Seen categories are indicated by the green colors while unseen categories by the blue colors.

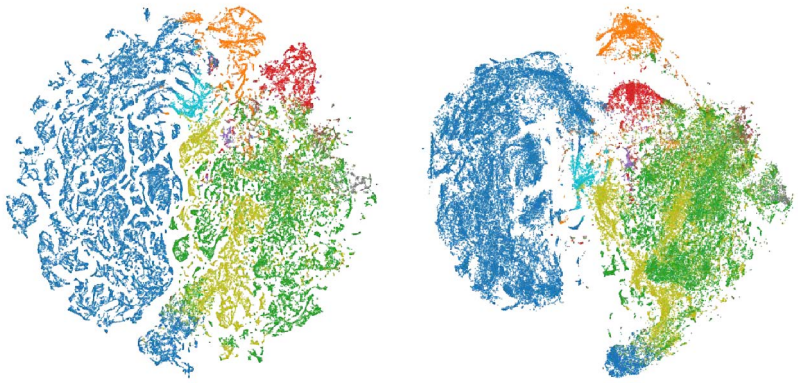
Table 4.5: Ablation analysis for the proposed modules of MPNet.

Exp	APG	AdaIN	CID	Seen Categories		Unseen Categories	
				1-shot	5-shot	1-shot	5-shot
a	-	-	-	58.5	60.6	25.4	28.3
b	✓	-	-	60.3	63.1	27.2	30.5
c	✓	✓	-	60.7	63.6	27.7	30.8
d	✓	-	✓	61.5	64.6	28.5	31.8
e	✓	✓	✓	**61.8**	**65.0**	**28.7**	**32.1**

among different pixel-level embedding feature. As shown in Table 4.5, we compare the class-aware average pooling (exp a) with our proposed APG (exp b). Our APG significantly promotes the performance by clear large margins on both seen categories (60.3 vs.58.5) and unseen ones (27.2 *vs.* 25.4). By introducing the AdaIN, the performance can be further enhanced to 60.7 and 27.7, respectively.

Contrastive Inter-Class Distraction In Table 4.5, we compare MPNet trained w/o CID (exp b) and w/ CID (exp d), to verify the effectiveness of the introducing inter-class discrepancy regularization. It can be seen that the CID makes consistent improvements on both seen categories (61.5 *vs.* 60.3) and unseen categories (28.5 *vs.* 27.2). To further verify the effectiveness of the CID in alleviating embedding ambiguity, we visualize the distribution of learned embedding features by t-SNE [171] as depicted in Fig 4.7. The feature embedding of w/ CID clearly shows a larger inter-category margin and higher intra-category compactness over that of w/o CID.

Visualization For additional qualitative evaluation, we illustrate results by our MPNet on *GFSP-Cityscapes* benchmark in Fig 4.6. We observe that even with only one annotated image as the guidance, MPNet outputs satisfactory parsing results. Due to the fact that our MPNet is never trained on the unseen categories, the parsing results of seen



(a) (b)

Figure 4.7: Embedding visualization with t-SNE of (a) w/o CID (b) w/ CID. Different colors indicate different categories.

categories are understandably better than the unseen ones.

4.5 Summary

Standing on the shoulders of former practice in object few-shot segmentation, this work steps further towards a more challenging yet general scenario, *i.e.*, generalized few-shot scene parsing. The success of this task will make the segmentation models well generalize to newly-emerging open visual concepts with little annotation labor. To this end, we present MPNet accordingly, in which the adaptive deep metric learning module and the contrastive inter-class distraction module are proposed to endow the learned meta parser with good generalization ability for complex scenarios. Although the MPNet achieves a preliminary success for the generalized few-shot scene parsing task, there is still a long way to go. We can observe that there is still a large performance gap between the seen and the unseen categories on the two benchmarks. Thus, more effective meta learning based algorithms are still required to alleviate this gap. We hope that our efforts will motivate more researchers and ease the future research on the generalized few-shot scene parsing task.

CHAPTER 5

CONSISTENT STRUCTURAL RELATION LEARNING FOR ZERO-SHOT SEGMENTATION

5.1 Preface

Semantic segmentation [137, 172] is a fundamental computer vision task that aims to assign a semantic label to each pixel in the given image. Although the development of FCN-based models [71, 138, 173] has significantly advanced semantic segmentation, the success of these approaches highly relies on cost-intensive and time-consuming dense mask annotations to train the network. To relieve the human effort in annotating accurate pixel-wise masks, there is an increasing interest in weakly-supervised segmentation and few-shot segmentation methods. Weakly supervised segmentation [63, 64] targets on learning segmentation models using lower-quality annotations such as image-level labels [174, 175], bounding boxes [176, 177] and scribbles [178, 179], which can be obtained more efficiently compared to pixel-wise masks. Meanwhile, few-shot segmentation [66, 68, 102, 169, 180] tackles the semantic segmentation from a meta-learning perspective and aims to perform segmentation with only a few annotated samples. Even significant progress has been made, these works are hard to completely liberate the request for mask annotations.

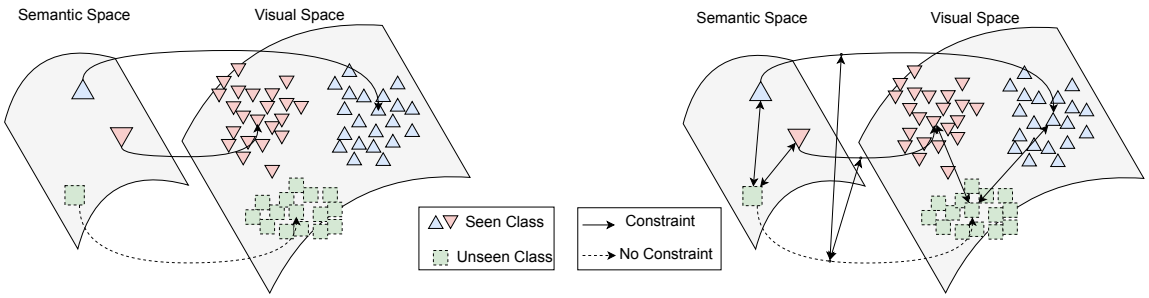
Most recently, Bucher *et al.* [22] took a step further to investigate how to effortlessly recognize those never-seen categories with zero training examples, and proposed a new learning paradigm, named Generalized Zero-Shot Semantic Segmentation (GZS3). Specif-

ically, during the training phase, in addition to the annotated images of seen categories, we are also provided with the semantic word embeddings of both seen and unseen labels. At test time, GZS3 aims to segment images containing pixels of all categories. As zero training examples of unseen categories are available, the key challenge of GZS3 lies in how to correctly recognize the pixels from these unseen categories. To tackle this, Bucher *et al.* [22] proposed a generative method by exploiting semantic word embeddings to generate unseen visual features, which are further employed to learn the classifiers for conducting segmentation. However, when training the generator from semantic space to visual space, they take each category independently with merely node-to-node knowledge transfer of seen categories. As shown in Figure 5.1a, no constraint is applied to guarantee the quality of generated visual features of unseen categories, resulting in poor generalization ability.

Hence, we seek to harness the inter-class relationship between seen and unseen categories to learn a better generator. We observe that different categories are roughly with similar relations in either semantic word embedding space or visual feature space. Therefore, we assume the relational structure embedded in the semantic space can be conveniently transferred to constrain the generated visual features of unseen categories. To this end, we propose Consistent Structural Relation Learning (CSRL) framework to tackle the challenging GZS3 task. Particularly, we propose a semantic-visual structural generator by integrating both feature generating and relation learning in a unified network architecture. Instead of taking each category independently, our CSRL generates the visual features from both seen and unseen categories, simultaneously. We additionally introduce the relational constraints from different structure granularities, including point-wise, pair-wise, and list-wise consistency, to facilitate the generalization of unseen categories. In this way, the learned visual features will be imposed to keep a consistent relational structure to their semantic-based counterparts, making the generator better adapt to unseen categories. Following [22], we conduct extensive experiments on two GZS3 benchmarks based on Pascal-VOC and Pascal-Context datasets. The proposed CSRL outperforms existing state-of-the-art methods by a large margin, resulting in ~7-12% on Pascal-VOC and ~2-5% on Pascal-Context.

5.2 Preliminaries

We denote a set of seen classes as $\mathscr{S}$ and a disjoint set of unseen classes as $\mathscr{U}$, where $\mathscr{S} \cap \mathscr{U} = \emptyset$. Let $\mathscr{D}_s = \{(\mathbf{x}, y | \mathbf{x} \in \mathscr{X}, y \in \mathscr{Y}^s\}$ represents the set of labeled training data



(a) Node-to-node generator (b) Structural generator

Figure 5.1: Illustration of CSRL. To achieve the goal of GZS3, we learn a generator to produce visual features from semantic word embeddings. Compared to (a) node-to-node generator, the proposed (b) structural generator explores the structural relations between seen and unseen categories to constrain the generation of unseen visual features.

on seen classes, where $\mathbf{x}$ is the pixel-wise feature embeddings from the visual space $\mathscr{X} \in \mathbb{R}^{d_v}$, y is the corresponding label in the label space $\mathscr{Y}^s$ of seen classes. Similar to the generalized zero-shot learning setting, in the task of GZS3, we aim to learn a model that takes an image as input and predicts the label of each pixel among both seen and unseen classes $\mathscr{S} \cup \mathscr{U}$. Clearly, without any side information, zero-shot learning is infeasible as there are no training samples of unseen classes. Thus, to achieve the goal of zero-shot learning, except the training set $\mathscr{D}_s$, we are also provided with the semantic word embeddings $\{\mathbf{a}_j | \mathbf{a}_j \in \mathscr{A}\}_{j=1}^{|\mathscr{S}\cup\mathscr{U}|}$ for both seen and unseen classes, where the semantic space $\mathscr{A} \in \mathbb{R}^{d_w}$. The d_w-dimensional semantic embeddings could be word representations (*e.g.*, word2vec [181] or GloVe embeddings [182]) or class attribute vectors [183]. In order to overcome the absence of unseen visual features, recent works [28, 29] adopt the generative model to produce unseen visual features. Specially, a generator $\mathscr{G} : \mathscr{A} \rightarrow \mathscr{X}$ is learned to generate visual features using corresponding word embeddings as input. Another benefit of these generative-based methods is that one can achieve the goal of zero-shot learning by directly adopting the existing CNN model (*e.g.*, Deeplab) without complex architecture modification. Concretely, the generator $\mathscr{G}$ is learned on seen classes and then generate visual features for unseen classes. A new classifier (usually the last layer of CNN) is retrained on real seen visual features and generated unseen visual features. At test time, the label of each pixel is predicted by selecting the category with the largest probability.

5.3 Consistent Structural Relation Learning for Zero-Shot Segmentation

As shown in Figure 5.2, we illustrate the details of the proposed CSRL framework. The goal of CSRL is to learn a better generator to produce visual features using semantic word embeddings as input. To achieve this goal, we introduce a semantic-visual structural generator to alternately update the node features of each category and the inter-category relations. We further exploit the structural relation consistency between seen and unseen categories to constrain the generating of unseen visual features. These structural relations include the point-wise, pair-wise and list-wise relations between seen and unseen categories. The generalized zero-shot semantic segmentation is achieved by learning on real seen visual features and the generated unseen visual features.

5.3.1 Semantic-Visual Structural Generator

Given a set of semantic word embeddings including samples from both seen and unseen categories, we aim to generate the corresponding set of synthetic visual features considering the relationships among categories. Such semantic-to-visual generation is achieved by a node-edge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, called semantic-visual structural generator in this work. The nodes $\mathcal{V} := \{\mathbf{v}_{i,n} | \forall i \in [1, |\mathcal{S} \cup \mathcal{U}|], n \in [1, N]\}$ in the graph denote the pixel-level feature embeddings with total N samples for category i. The edges $\mathcal{E} := \{e_{ij} | \forall i, j \in [1, |\mathcal{S} \cup \mathcal{U}|]\}$ are constructed based on the relationships between prototypes of category i and j.

The structural generator consists of L layers, where each layer contains a feature aggregation step to update the node feature and a relation aggregation step to update the edge feature. We denote $\mathbf{v}_i^\ell$ and e_{ij}^ℓ as the node feature and the edge feature of layer $\ell \in [1, L]$, respectively.

As the semantic word embedding $\mathbf{a}_i$ is a deterministic value, we enhance the feature diversity by concatenating a random variable $\mathbf{z}$ with a Gaussian distribution. Thus, node features are initialized by the semantic word embeddings $\mathbf{v}_{i,n}^0 = [\mathbf{a}_i \oplus \mathbf{z}_{i,n}]$, where $\oplus$ denotes the concatenation operation. Edge features $e_{ij}^0 = \mathbf{a_i} \cdot \mathbf{a_j} / \|\mathbf{a_i}\|_2 \|\mathbf{a_j}\|_2$ are initialized by the cosine similarity between semantic word embeddings.

Feature Aggregation To alleviate the issue introduced by abnormal samples, especially only a limited number of samples in one categories, we aggregate the feature representation based on the category prototypes instead of raw samples. Specially, the

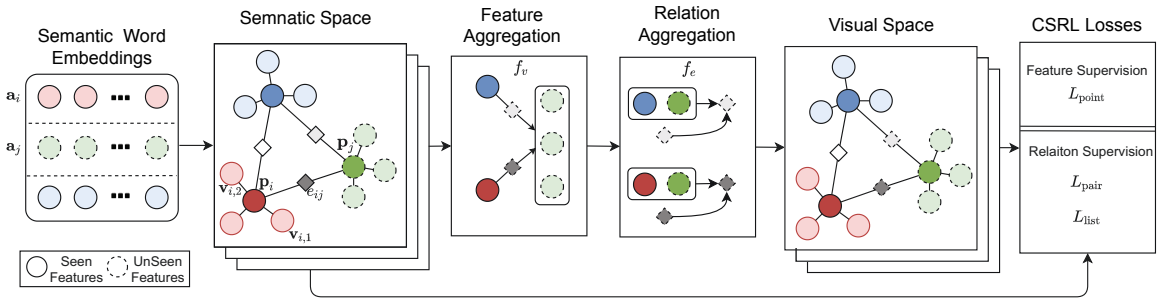


Figure 5.2: The framework of the proposed CSRL. Our CSRL incorporates the feature generating and relation learning into a unified architecture. Given the semantic word embedding, CSRL generates visual features by alternately feature and relation aggregation. The proposed CSRL is trained under supervision from point-wise consistency on seen classes, pair-wise and list-wise consistency across seen and unseen classes.

category prototype $\mathbf{p}_i$ is defined as,

$$\mathbf{p}_i^{\ell-1} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{v}_{i,n}^{\ell-1}. \tag{5.1}$$

After calculating all prototype representations $\{\mathbf{p}_i|\forall i \in [1,|\mathscr{S}\cup\mathscr{U}|]\}$, we are able to propagate the relevant knowledge from other categories based on the edge features. The node feature aggregation of the l-th layer follows,

$$\mathbf{v}_{i,n}^{\ell} = f_v^{\ell}([\mathbf{v}_{i,n}^{\ell-1} \oplus \sum_{j=1,j\neq i}^{|\mathscr{S}\cup\mathscr{U}|} e_{ij}^{\ell-1}\mathbf{p}_i^{\ell-1}];\phi_v^{\ell}). \tag{5.2}$$

where f_v is a transformation network with parameters ϕ_v^{ℓ}.

Relation Aggregation After aggregate the node features, the edge feature aggregation is processed based on the newly updated node features, The edge feature aggregation of the l-th layer follows,

$$e_{ij}^{\ell} = f_e^{\ell}(|\mathbf{p}_i^{\ell} - \mathbf{p}_j^{\ell}|;\phi_e^{\ell})e_{ij}^{\ell-1}, \tag{5.3}$$

where f_e is a transformation network with parameters ϕ_e^{ℓ}.

By alternately feature aggregation and the relation aggregation steps, we simultaneous achieve the feature generating and relation learning. At $\ell = L$, the output nodes are the generated visual features $\hat{\mathbf{x}}$ including both seen and unseen categories, while the edge features are the learned relations between categories.

5.3.2 Consistent Structural Relation Learning

The key to generalized zero-shot segmentation is the ability to generate visual features $\hat{\mathbf{x}} \in \hat{\mathscr{X}}$ conditioned on the semantic word embedding $\mathbf{a}$, even without access to any image pixels of this category. In order to learn a better generator, we explore the relation constraints from different structure granularities as supervision signals to train the generator $\mathscr{G}$.

Point-wise consistency At training time, only the real visual features from seen categories are available to access. Thus, on these seen categories, we optimize the distribution divergence between real visual features and generated visual features as supervision signals. As this divergence reflects the consistency of every single category between real and generated visual feature distributions, here we note it as *point-wise consistency*. Here, we minimize distribution divergence on seen categories by optimizing the *maximum mean discrepancy* as,

$$\mathscr{L}_{\text{point}} = \frac{1}{|\mathscr{S}|}\sum_{c=1}^{|\mathscr{S}|}[\mathbb{E}_{\mathbf{x},\mathbf{x}'\sim\mathscr{X}^c}K(\mathbf{x},\mathbf{x}') + \mathbb{E}_{\hat{\mathbf{x}},\hat{\mathbf{x}}'\sim\hat{\mathscr{X}}^c}K(\hat{\mathbf{x}},\hat{\mathbf{x}}') - 2\mathbb{E}_{\mathbf{x}\sim\mathscr{X}^c,\hat{\mathbf{x}}\sim\hat{\mathscr{X}}^c}K(\mathbf{x},\hat{\mathbf{x}})], \tag{5.4}$$

where K is the Gaussian kernel with bandwidth parameter σ defined as $K(\mathbf{x},\mathbf{x}') = \exp(-\frac{1}{2\sigma^2}\|\mathbf{x}-\mathbf{x}'\|^2)$.

By optimizing the point-wise consistency on seen categories, there is no explicit constraint on the generation of unseen categories. Thus the quality of produced unseen features purely relies on the generalization ability of the generator. To enhance and constrain the visual feature generation especially on unseen categories, we transfer the structural relations on semantic word embedding space to the generator visual features space. In this section, we consider the *pair-wise consistency* and *list-wise consistency*. The pair-wise relations reflect that the feature similarity between two categories, *i.e.*, one seen category and one unseen one, should be consistent on both semantic space and visual space. The list-wise relations require that the relation ranking permutation order should also be consistent on semantic space and visual space.

Pair-wise consistency We extract the relation matrix between unseen and seen categories from the edge features in structural generator $\mathscr{G}$ as $\mathbf{M} = \{e^{\ell}_{ij} | \forall i \in [1, |\mathscr{U}|], j \in [1, |\mathscr{S}|]\} \in \mathbb{R}^{|\mathscr{U}|\times|\mathscr{S}|}$. For each unseen category, the relation values is further normalized by applying softmax function as follows,

$$\tilde{e}^{\ell}_{ij} = \frac{\exp(e_{ij}/\gamma)}{\sum_{j'=1}^{|\mathscr{S}|}\exp(e_{ij'}/\gamma)}, \tag{5.5}$$

where γ is a scaling factor to soften the relation distribution. Thus, in the semantic word embedding space (*i.e.*, the input layer $\ell = 0$), we have the relation matrix as $\mathbf{M}^{\mathscr{A}}$. In the generated visual feature space (*i.e.*, the output layer $\ell = L$), the relation matrix is denote as $\mathbf{M}^{\hat{\mathscr{X}}}$.

To maintain the pair-wise relation consistency between semantic space and visual feature space, we adopt the Kullback-Leibler divergence as the learning objective. Concretely, the *pair-wise consistency* is defined as,

$$\mathscr{L}_{\text{pair}}(\mathbf{M}^{\mathscr{A}}, \mathbf{M}^{\hat{\mathscr{X}}}) = \frac{1}{|\mathscr{U}|} \sum_{i=1}^{|\mathscr{U}|} D_{\text{KL}}[\mathbf{M}_i^{\mathscr{A}} || \mathbf{M}_i^{\hat{\mathscr{X}}}]. \tag{5.6}$$

List-wise consistency Instead of only focus on the relationship from a pair of categories at a time, inspired by [184, 185], we further investigate the entire ranking permutation of the relation list as complementary supervision. The core idea is that we take the relation ranking as a distribution rather than a deterministic order. We aim to associate the probability with every rank permutation between semantic space and visual space. Given one permutation π of the relation list, where $\pi(i)$ denotes the i-th list index of this permeation. We calculate the probability of this ranking permutation as,

$$P(\pi|\mathbf{M}_i) = \prod_{j=1}^{|\mathscr{S}|} \frac{\exp(e_{i\pi(j)}/\gamma)}{\sum_{k=j}^{|\mathscr{S}|} \exp(e_{i\pi(k)}/\gamma)} \tag{5.7}$$

where γ is a scaling factor.

We aim to maintain all possible relation ranking permutations $\pi \in \mathscr{P}$ as consistent as possible both on semantic space and visual features space. Similar to pair-wise consistency, the *list-wise consistency* is defined as,

$$\mathscr{L}_{\text{list}}(\mathbf{M}^{\mathscr{A}}, \mathbf{M}^{\hat{\mathscr{X}}}) = \frac{1}{|\mathscr{U}|} \sum_{i=1}^{|\mathscr{U}|} D_{\text{KL}}[P(\pi \in \mathscr{P}|\mathbf{M}_i^{\mathscr{A}}) \| P(\pi \in \mathscr{P}|\mathbf{M}_i^{\hat{\mathscr{X}}})] \tag{5.8}$$

5.3.3 Training and Inference

In this subsection, we introduce the whole procedures to achieve GZS3. During the training stage, we start from training an off-the-shelf segmentation model (*e.g.*, DeepLabv3+) on all annotated data from seen categories. After training on seen categories, we remove the last classification layer and the remaining network serves as a visual features extractor to get the training set of seen categories $\mathscr{D}_s$. Then, we train our semantic-visual structural generator $\mathscr{G}$ under the supervision of consistent structural relation learning losses,

$$\mathscr{L}(\phi) = \mathscr{L}_{\text{point}} + \mathscr{L}_{\text{pair}} + \mathscr{L}_{\text{list}}. \tag{5.9}$$

Table 5.1: Generalized zero-shot semantic segmentation performance on Pascal-VOC dataset.

Settings	Methods	Seen mIoU	Unseen mIoU	Overall mIoU	Overall hIoU
unseen-2	SegDevis	68.1%	3.2%	44.1%	6.1%
	SPNet	71.8%	34.7%	68.2%	46.8%
	ZS3Net	72.0%	35.4%	68.5%	47.5%
	CSRL	**73.4%**	**45.7%**	**70.7%**	**56.3%**
unseen-4	SegDevis	64.3%	2.9%	38.9%	5.5%
	SPNet	67.3%	21.8%	58.6%	32.9%
	ZS3Net	66.4%	23.2%	58.2%	34.4%
	CSRL	**69.8%**	**31.7%**	**62.5%**	**43.6%**
unseen-6	SegDevis	39.8%	2.7%	33.4%	5.1%
	SPNet	64.5%	20.1%	51.8%	30.6%
	ZS3Net	47.3%	24.2%	40.7%	32.0%
	CSRL	**66.2%**	**29.4%**	**55.6%**	**40.7%**
unseen-8	SegDevis	35.7%	2.0%	24.3%	3.8%
	SPNet	61.2%	19.9%	45.5%	30.0%
	ZS3Net	29.2%	22.9%	26.8%	25.7%
	CSRL	**62.4%**	**26.9%**	**48.8%**	**37.6%**
unseen-10	SegDevis	31.7%	1.9%	16.9%	3.6%
	SPNet	59.0%	18.1%	39.5%	27.7%
	ZS3Net	33.9%	18.1%	26.3%	23.6%
	CSRL	**59.2%**	**21.0%**	**50.0%**	**31.0%**

To maintain simplicity, here we directly add these three terms. Once the generator $\mathscr{G}$ is trained, arbitrarily many visual features can be generated from semantic word embeddings, especially for unseen categories. In this way, we build a generated unseen training set denote as $\hat{\mathscr{D}}_u = \{\hat{\mathbf{x}}, y | \hat{\mathbf{x}} \in \hat{\mathscr{X}}, y \in \mathscr{Y}^u\}$. A new pixel-level classifier is trained on the combined training set including real seen visual features from $\mathscr{D}_s$ and generated unseen visual features from $\hat{\mathscr{D}}_u$. In this way, the new model can be used to conduct generalized zero-shot semantic segmentation of a given image that exhibit categories from both seen and unseen classes.

5.4 Experiments

5.4.1 Experiment Settings

Datasets We conduct experiments on two datasets including Pascal-VOC [76] and Pascal-Context [82]. Pascal-VOC focuses on object semantic segmentation scenario, which contains 10,582 training and 1,449 validation images from 20 classes. Pascal-Context targets on the scene parsing scenario, which comprises 4,998 training and 5,105 validation images from 59 classes. Following [22], we construct zero-shot segmentation

setups with different number of unseen classes, including 2, 4, 6, 8 and 10 unseen classes, and all the rest ones are the seen classes. Concretely, the unseen class set is extended in an incremental manner, *i.e.*, the 4-unseen set contains the 2-unseen set. The unseen class splits are *2-cow / motorbike, 4-airplane / sofa, 6-cat / tv, 8-train / bottle, 10-chair / potted-plant* for Pascal-VOC dataset and *2-cow / motorbike, 4-sofa / cat, 6-boat / fence, 8-bird / tvmonitor, 10-keyboard / aeroplane* for Pascal-Context dataset.

Evaluation Metrics In our experiments, similar to the standard semantic segmentation task, we adopt mean intersection-over-union (mIoU) as the principal metric. The generalized zero-shot semantic segmentation focuses on the overall performance including both seen and unseen categories. To avoid the performance on seen categories dominates, we also report the harmonic mean (hIoU) of seen mIoU and unseen mIoU suggested by [186],

$$hIoU = \frac{2 * mIoU_s * mIoU_u}{mIoU_s + mIoU_u}. \tag{5.10}$$

Implementation Details We choose the DeeplabV3+ [71] with ResNet-101 [5] as our segmentation network. The ImageNet [156] covers a wide range of categories, where most unseen categories are actually included. Therefore, directly adopting the publicly ImageNet pre-trained model may break the setting of zero-shot learning. To avoid the supervision leakage from unseen classes, we employ the model provided by [22], which is solely pre-trained using seen categories. For the aggregation network f_e and f_v in Sec 5.3.1, we use the multi-layer perception network proposed by [187]. We implemented our method both by the Pytorch platform and the PaddlePaddle platform, both achieving similar performance.

5.4.2 Comparisons with State-of-the-art Methods

We compare our proposed CSRL with SegDeViSe [188], SPNet [21], ZS3Net [22]. SegDeViSe regresses semantic word features from pixel-level visual features, which is learned by maximizing the cosine similarity between the output and the target word embeddings. SPNet encodes images in the word embedding space and uses a semantic projection layer to produce class probabilities. ZS3Net is the current state-of-the-art method, which generates unseen visual features from word embeddings to achieve zero-shot segmentation. All these methods adopt the same segmentation network, *i.e.*, DeepLabV3+, for a fair comparison. The key commonality shared by these methods is: they take each category as an independent point without considering its relations to other categories. Differently,

Table 5.2: Generalized zero-shot semantic segmentation result on Pascal-Context dataset.

Settings	Methods	Seen mIoU	Unseen mIoU	Overall mIoU	Overall hIoU
unseen-2	SegDevis	35.8%	2.7%	33.1%	5.0%
	SPNet	38.2%	16.7%	37.5%	23.2%
	ZS3Net	41.6%	21.6%	41.0%	28.4%
	CSRL	**41.9%**	**27.8%**	**41.4%**	**33.4%**
unseen-4	SegDevis	33.4%	2.5%	30.7%	4.7%
	SPNet	36.3%	18.1%	35.1%	24.2%
	ZS3Net	37.2%	**24.9%**	36.4%	29.8%
	CSRL	**39.8%**	23.9%	**38.7%**	**29.9%**
unseen-6	SegDevis	31.9%	2.1%	28.8%	3.9%
	SPNet	31.9%	19.9%	30.7%	24.5%
	ZS3Net	32.1%	20.7%	30.9%	25.2%
	CSRL	**35.5%**	**22.0%**	**34.1%**	**27.2%**
unseen-8	SegDevis	22.0%	1.7%	19.2%	3.2%
	SPNet	28.6%	14.3%	26.7%	19.1%
	ZS3Net	20.9%	16.0%	20.3%	18.1%
	CSRL	**31.7%**	**18.1%**	**29.9%**	**23.0%**
unseen-10	SegDevis	17.5%	1.3%	14.3%	2.4%
	SPNet	27.1%	9.8%	24.3%	14.4%
	ZS3Net	20.8%	12.7%	19.4%	15.8%
	CSRL	**29.4%**	**14.6%**	**27.0%**	**19.5%**

we generate the unseen visual features by exploring the structural relations between categories.

We report the performance of generalized zero-shot semantic segmentation on Pascal-VOC dataset in Table 5.1 and Pascal-Contex dataset in Table 5.2. Results of SPNet are based on our implementation, and other results of ZS3Net and SegDeVis are directly taken from paper [22]. In these two tables, first, we observe that the generative methods (*i.e.*, ZS3Net, CSRL) significantly outperforms semantic embedding-based methods (*i.e.*, SegDeViSe, SPNet). The semantic embedding-based methods, although perform well on seen categories, achieve a large performance drop for unseen ones. By leveraging structural relation consistency to better guide the generation of unseen visual features, our CSRL provides significant gains particularly on the unseen classes (*e.g.*, +10.3% for the 2-unseen split in terms of unseen mIoU). Second, our CSRL significantly outperforms others by large margins for various splits (~7-12% for hIoU), which can well demonstrate the effectiveness of the consistent structural relation learning framework. Third, our CSRL also achieves large performance gains on the more challenging benchmark Pascal-Context, which requires densely predictions for the full images. The qualitative comparison between ZS3Net and CSRL is shown in Figure 5.3. We can observe that our CSRL achieves much better segmentation results and successfully recognize the unseen

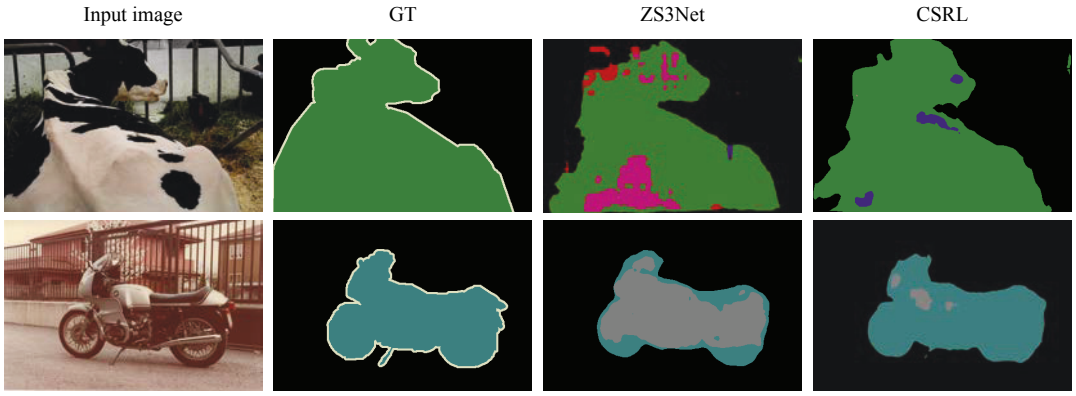


Figure 5.3: Qualitative comparisons on Pascal-VOC dataset under the unseen-2 setting.

objects (*e.g.cow* and *motorbike*) where the ZS3Net mostly fails.

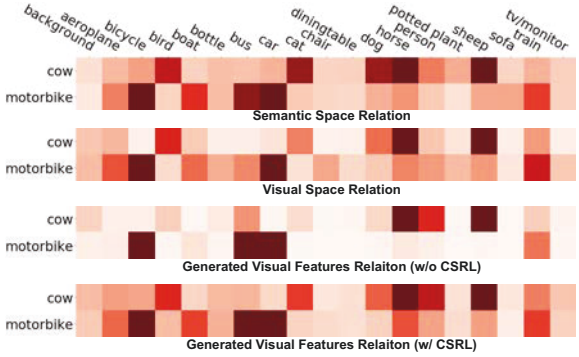


Figure 5.4: Relations between unseen (cow and motorbike) and seen categories.

Table 5.3: Ablation study of CSRL on Pascal-VOC.

Exp	Point	Pair	List	Seen mIoU	Unseen mIoU	Overall mIoU	Overall hIoU
I	✓	-	-	73.0%	40.3%	69.8%	51.9%
II	✓	✓	-	73.4%	43.3%	70.5%	54.5%
III	✓	-	✓	73.0%	42.7%	70.1%	53.9%
CSRL	✓	✓	✓	**73.4%**	**45.7%**	**70.7%**	**56.3%**

5.4.3 Ablation Analysis

Quantitative analysis for structural relations We conduct extensive quantitative analysis for those key components in CSRL. In Table 5.3, we compare the effects of different structural relations with the unseen-2 split on Pascal-VOC. First, simply performing node-to-note generation results in the hIoU score of 51.9% (I). Second, by introducing the pair-wise relation for optimization, the hIoU will be significantly enhanced by 3.6% (II). Third, by replacing pair-wise relation with list-wise relation, the improvement is still notable, *i.e.*, 2.0% (III). Finally, simultaneously considering all the components will lead to the best hIoU score of 56.3% (CSRL).

Qualitative analysis for inter-category relations In Figure 5.4, we visualize the inter-category relations between unseen categories (*i.e.*, cow and motorbike) and seen categories based on different feature embeddings. The relations are normalized for better visualization. The darker the colors, the stronger the relations. “Semantic Space Relation" and “Visual Space Relation" indicate cosine similarities of using word2vec features and CNN features with supervised training, respectively. First, we can observe that semantic relations between unseen and seen categories keep consistent across different feature spaces. Second, by introducing the CSRL, the relations of generated visual features will be more consistent compared to those without CSRL, leading to better discriminative ability.

5.5 Summary

In this section, to tackle the challenging generalized zero-shot semantic segmentation task, we proposed a simple yet effective framework called Consistent Structural Relation Learning (CSRL). We propose a semantic-visual structural generator by integrating both feature generating and relation learning in a unified network architecture. We effectively explore relation consistency from multiple structure granularities to better guide the generation of unseen visual features. The proposed CSRL achieves the new state-of-the-art on two zero-shot segmentation benchmarks, which outperforming the former practices by a large margin. Although CSRL achieves a large improvement for the generalized zero-shot semantic segmentation, there is still a long way to go. We can observe that there is still a large performance gap between the seen and the unseen categories on the two benchmarks. Thus, more effective GZS3 algorithms are still required to alleviate this gap. We hope that our efforts will motivate more researchers and ease future research.

CHAPTER 6

SUPER-RESOLVING CROSS-DOMAIN FACE MINIATURES BY PEEKING AT ONE-SHOT EXEMPLAR

6.1 Preface

Face Super-Resolution (FSR), also known as face hallucination, aims at reconstructing high-resolution (HR) face images from input low-resolution (LR) ones. FSR provides critical information for the downstream computer vision and machine learning tasks, such as face detection [2], recognition [189] and photo-editing [190–192]. Thanks to the advance of generative adversarial networks [26], FSR has achieved great success in recent years [116, 117, 121, 124, 125, 127, 193–196].

Previous FSR methods usually presume training and testing LR faces are captured from the same domain. When testing LR faces resemble the training ones, previous works achieve authentic upsampled HR faces. However, in practice, the domain gap between testing images and training ones is inevitable due to different imaging equipment, illumination conditions, *etc*. As shown in the upper right of Figure 6.1, previous state-of-the-art FSR methods fail to upsample HR authentically due to the large domain gap between the target domain (testing) and source domain (training). Considering FSR models would be deployed in different scenarios, it is very inefficient to re-train every deployed FSR model by collecting large-scale data from the corresponding target domain. Therefore, only using a few samples, ideally one example, to efficiently update an FSR model is highly desirable.

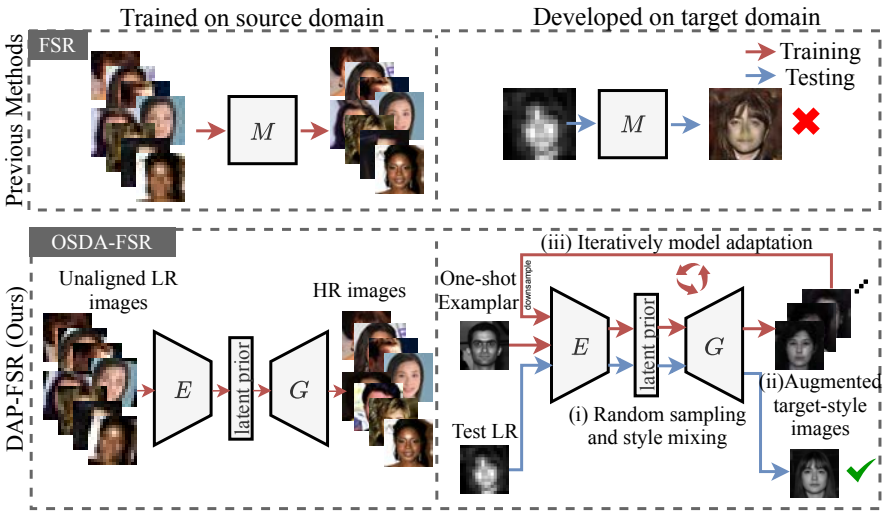


Figure 6.1: Conventional FSR methods achieve good performance on the source dataset, but are prone to fail on the target dataset due to the domain gap. Our proposed method effectively adapts the model by leveraging only one-shot example.

In this work, we aim to super-resolve LR faces that exhibit an obvious domain gap by only leveraging one-shot exemplar from the target domain. We name this task as One-Shot Domain Adaption for Face Super-Resolution (OSDA-FSR). Different from conventional FSR methods [116, 121, 122], two challenges are naturally raised: (i) how to design a FSR network architecture that is intrinsically suitable for efficient adaptation; and (ii) how to explore one example to bridge the domain gap since simply fine-tuning an FSR network with one example is ineffective.

To address these challenges, we present a novel Domain-Aware Pyramid based Face Super-Resolution network, namely DAP-FSR network. Our DAP-FSR contains two parts: a domain-aware pyramid encoder and an upsampling decoder. Our DAP-FSR encoder is designed to extract the latent representations by leveraging the multi-scale features from the input LR faces. Considering LR faces may be unaligned, we propose an Instance Spatial Transformer Networks (ISTN) to align LR faces inspired by [197]. In this way, we facilitate the latent representation learning and face upsampling processes by aligning the LR faces into the canonical view. Motivated by the powerful architecture of StyleGAN [198, 199], an image generation network, we construct our upsampling decoder. Once we obtain the latent representations, we feed those representations to our DAP-FSR decoder to hallucinate high-quality HR face images.

To tackle the problem of super-resolving LR faces in a new domain without the need for tremendous data collection, we propose a Domain-Aware latent Mixing and

Model Adaptation algorithm (DAMMA). In a nutshell, our DAMMA algorithm is able to adapt the model trained on the source domain to the target domain by exploring only the one-shot example. As illustrated in Figure 6.1, when a target domain example is given, DAP-FSR network first extracts its latent representations. Then, supervised by the given one-shot example, we learn a soft mixture weight to mix the target latent representations with random-sampled source latent ones. In this fashion, the newly generated faces will resemble the target domain faces and we significantly augment the target-style data. By constrained fine-tuning our decoder with the augmented images, our network is gradually adapted from the source domain to the target domain. After iteratively updating the soft mixing weight and adapting our decoder, our DAP-FSR attains authentic target domain HR faces.

Our main contributions are summarized as follows,

- We propose a novel domain-aware pyramid-based face super-resolution network, named DAP-FSR network, to efficiently upsample cross-domain LR face images by peeking at one-shot target domain example.

- We present a simple yet effective domain-aware latent mixing and model adaptation algorithm (DAMMA) to adapt our DAP-FSR to the target domain. Our DAMMA generates target-style alike faces to adapt the upsampling decoder in DAP-FSR by fully exploiting the one-shot example.

- To the best of our knowledge, our method is the first attempt to super-resolve cross-domain LR face images, making our method more practical.

- Our proposed DAP-FSR can be adapted to a target domain effectively and is also robust to unaligned LR faces. Experiments on three constructed cross-domain face super-resolution benchmarks validate the superior performance of our proposed approach compared to the state-of-the-art methods.

6.2 Task Definition: One-shot based FSR

Conventional Face Super-Resolution (FSR) methods aim to learn a face super-resolution model M that generates a high-resolution super-resolved face image $I_{SR} \in \mathbb{R}^{H \times W}$ from a low-resolution one $I_{LR} \in \mathbb{R}^{h \times w}$, as follows:

$$I_{SR} = M(I_{LR}). \tag{6.1}$$

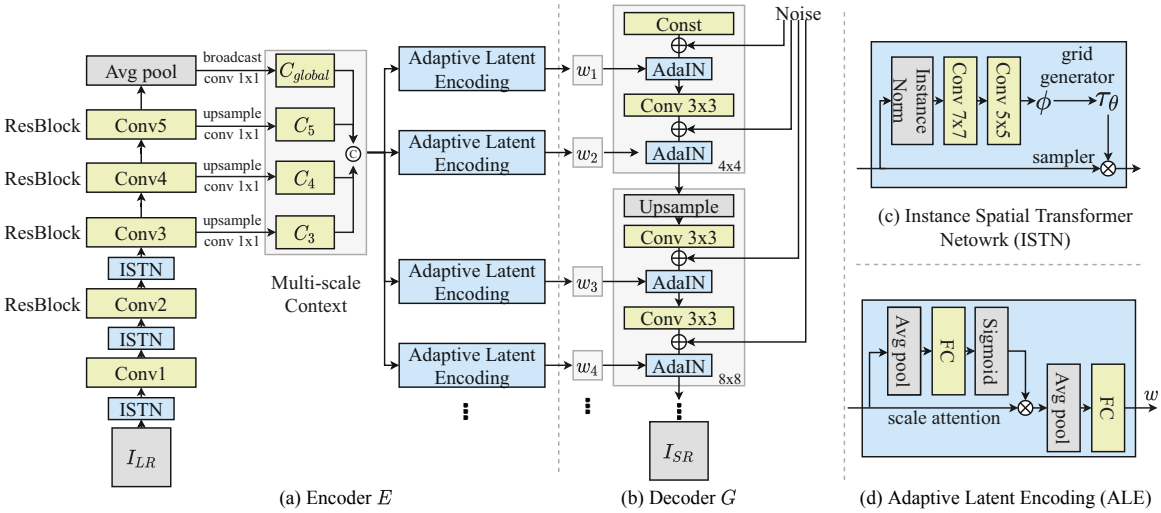


Figure 6.2: Illustration of our DAP-FSR architecture. (a) The encoder network. Feature maps from different spatial resolution are up-sampled and concatenated as the multi-scale pyramid context. Each Adaptive Latent Encoding (ALE) module dynamically attends the multi-scale context to generate the latent representation $\mathbf{w}_i$. (b) The decoder network, where the HR images are generated based on the latent representations. (c) The Instance Spatial Transformer Network (ISTN) learns the style-invariant affine transformation matrix to adjust the unaligned LR images. (d) The detailed Adaptive Latent Encoding module, where the channel-wise feature attention is learned to adaptively capture the multi-scale information of the input images.

The goal of the FSR task is to make the reconstructed image I_{SR} best recover its corresponding high-resolution version I_{HR}. In the conventional face hallucination setting [110, 112, 116], an FSR model M is trained and evaluated on the $\{(I_{LR}, I_{HR})\}$ pairs from the same source domain. However, as illustrated in Figure 6.1, when LR images come from another target domain, a pre-trained model M might fail to generalize well to the new domain data and the quality of super-resolved HR images will degrade severely.

Inspired by previous domain adaptation works [200], we formulate our task as One-Shot Domain Adaptation for Face Super-Resolution (OSDA-FSR). In general, OSDA-FSR can be divided into two stages, *i.e.*, a procurement stage and a deployment stage, based on the real-world application scenario. In the procurement stage, an FSR model is trained on the large-scale source dataset with N_s HR and LR image pairs, denoted as $\mathscr{D}_s = \{(I^s_{LR}, I^s_{HR})_i\}_{i=1}^{N_s}$. In this stage, image reconstruction objectives will be employed to optimize the model parameters. However, in the deployment stage, a trained model might encounter an unknown data distribution shift in a target domain. In this case, a deep model may fail to super-resolve LR faces in a target domain without knowing any

information about the new domain.

Although collecting data and re-training a network can solve this issue, it might be inefficient and time-consuming when deploying deep models in many different real-world scenarios. Therefore, we aim to use only a few examples, *e.g.*, K LR-HR pairs $\mathscr{D}_t = \{(I^t_{LR}, I^t_{HR})_i\}^K_{i=1}$, to effectively adapt the pre-trained model M. Without the loss of generality, we focus on the most challenging case where $K = 1$. In other words, we will exploit the one-shot exemplar to minimize the domain gap and then hallucinate the target domain LR faces.

6.3 Proposed Method

Overview The general goal of OSDA-FSR task is to transfer the model from the trained source domain to the target domain by fully exploiting the given one-shot example. To achieve this goal, the key idea of our approach is to adapt the model towards the target domain by enriching the target-style samples beyond the solely given one-shot exemplar. We present a Domain-Aware Pyramid-based Face Super-Resolution (DAP-FSR) network to super-resolve input LR images to output HR images, as shown in Figure 6.2. Our DAP-FSR firstly obtains the semantic latent representations from an unaligned LR face image by the encoder network and then generates the high-quality HR images from these latent representations by the upsampling decoder network.

Given an LR image in the target domain, our DAP-FSR network first extracts the latent representations. However, due to the existing large domain gap, the latent representations of target domain LR images may not lie on the manifold of the source domain ones, thus causing inferior upsampled results. To address this problem, we propose to project the latent representations of the target one-shot exemplar to the closest one in the source domain. We then synthesize random images sharing similar styles with the target domain by mixing randomly sampled source and the extracted target domain latent representations. These generated samples will be in turn used to optimize our upsampling network. In this fashion, the latent representation manifold will gradually shift to the target domain and we can super-resolve target domain LR images even with only one exemplar.

6.3.1 Domain Aware Pyramid-based FSR

Choice of decoder and latent space Due to the advanced network architecture, StyleGAN [198, 199] obtains phenomenal high-resolution and photo-realistic images. Recent work [122] also demonstrates the possibility that employing a pre-trained StyleGAN, HR faces can be found from the given LR inputs. More importantly, decouple the training of an encoder and a decoder would allow us to achieve larger upscaling factors while being less restricted by GPU memory. Therefore, we choose the StyleGAN architecture as the upsampling decoder in our DAP-FSR.

Former work [122] demonstrates that the multi-layer disentangled latent space $\mathcal{W}+$ in StyleGAN is more representative to depict an image than the normalized Gaussian distribution space $\mathcal{Z}$. Furthermore, the layer-wise corresponding AdaIN modules in StyleGAN can also facilitate us to transfer domain-specific characteristics when we adapt our trained upsampling decoder to a target domain. Hence, to fully utilize the power of StyleGAN, we adopt the $\mathbf{w} \in \mathbb{R}^{l \times d_w}$ as our latent representations to better encode the LR images, where l is the layer number and d_w is the latent representation dimension.

Latent representation learning Unlike PLUSE [122] that optimizes a latent representation $\mathbf{w} \in \mathcal{W}+$ by minimizing the pixel-wise reconstruction loss between the downsampled version of upsampled HR image and the input image, we introduce an encoder to extract latent representations of the input LR faces. Doing so allows us to address unaligned LR faces and handle the domain gap by fine-tuning our upsampling decoder, while PLUSE cannot handle the domain gap and face misalignments as its decoder (*i.e.*, pre-trained StyleGAN) is fixed and only $\mathbf{w}$ is updated during iterations.

Recall that in the StyleGAN, each latent representation controls a certain level of image details. Hence, our encoder aims to adaptively predict latent representations from an enhanced multi-scale context feature. Toward this goal, we develop an adaptive latent encoding (ALE) module that is able to generate latent representations for the upsampling decoder at different scales adaptively. Here, we employ ResNet50 as our encoder to extract multi-scale feature maps at the `conv3`, `conv4`, `conv5` and average pooling layers, denoted as $C_3, C_4, C_5, C_{global}$, as shown in Figure 6.2. Then, each ALE generates multi-scale latent representations $\mathbf{w}_i$ for the decoder by attending the multi-scale features adaptively. Then, the latent representations are fed to our upsampling decoder for face hallucination.

Robust against unaligned LR faces Previous face hallucination methods [112, 122, 189] often assume LR faces are precisely aligned beforehand. However, such an

Algorithm 2: Domain-Aware Latent Mixing and Model Adaptation

Input: Initialized DAP-FSR model $M = (E, G)$ trained on source dataset $\mathscr{D}_s$, one-shot exemplar $\{I^t_{LR}, I^t_{HR}\} \in \mathscr{D}_t$, initialized latent code mixing weight α_0, AdaIN parameter ϕ in G ,learning rate ξ, η

Output: Adapted model M_{ϕ^*}

while *do not converge* **do**

  Generate $\mathbf{w}^t$ by manifold preserving projection as equation 6.2;

  Sample a batch of source latent codes: $\mathbf{w}_s = \mu_{\mathbf{w}} + \sigma_{\mathbf{w}}\epsilon, \epsilon \sim \mathcal{N}(0,1)$;

  Initialize latent code mixing weight: $\alpha \leftarrow \alpha_0$;

  for *i=1,2,3,...,n* **do**

    Update mixing weight by equation 6.7: $\alpha \leftarrow \alpha - \xi \nabla_\alpha \mathscr{L}(\alpha)$;

    Generate mixing latent codes $\mathbf{w}^m$ by equation 6.6;

    Update model parameters by equation 6.5: $\phi \leftarrow \phi - \eta \nabla_\phi \mathscr{L}(\phi)$;

  end

end

Return final model weight ϕ as ϕ^*;

assumption hardly holds in real application scenarios. Inspired by the works [193, 196], we estimate the transformation of LR images and warp them to the canonical position by the spatial transformation network (STN) [197]. Therefore, our network is robust against unaligned LR faces with in-plane rotations, translations and scale changes. The detailed architecture of spatial transformation layers are illustrated in Figure 6.2(c).

More importantly, unlike previous FSR models [193, 196] that use STNs, we apply an instance normalization layer to the feature maps before computing the transformation parameters in our instance spatial transformer network (ISTN) module. This allows us to obtain style-invariant feature maps. Therefore, even when target-domain LR faces are provided, our ISTN layers are still able to align them to the up-right position, potentially facilitating the following domain adaptation process. Thus, our decoder can focus on super-resolving high-quality HR faces while preserving the latent representations from being affected by misaligned input LR faces.

Manifold preserving encoding Previous work [201] shows that it is possible to invert an arbitrary image, even not a face image, into style latent space $\mathscr{W}+$. However, such deduced latent codes are not aligned with the semantic knowledge prior learned by $G(\cdot)$ and lose the versatile image editing capability. In our OSDA-FSR task, the situation will become even worse where a domain gap between the source and target domains exists.

To overcome these drawbacks, we explicitly constrain the output of the encoder $E(\cdot)$ in the feature space of G. Particularly, instead of directly predicting the style latent codes, we predict the offset scale *w.r.t.* the mean $\mu_{\mathbf{w}}$ and variance $\sigma_{\mathbf{w}}$ of the

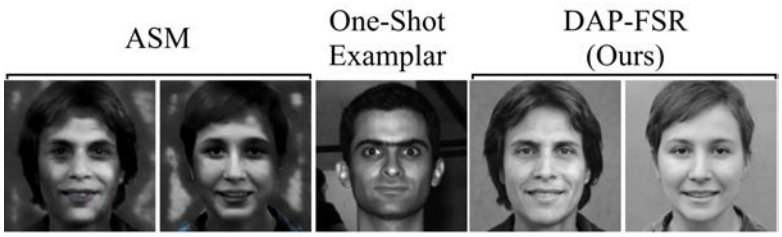


Figure 6.3: Compared to the style-transfer based method ASM [1] (left), given only one-shot target domain exemplar (ExtendedYaleB), our method (right) efficiently generates authentic target-style images from the source domain (CelebA).

latent representations of G. To be specific, our DAP-FSR model maps the encoded representations of LR images to the latent representations $\mathbf{w}$ of the decoder, as follows:

$$\mathbf{w} = \mu_{\mathbf{w}} + E(I_{LR})\sigma_{\mathbf{w}}, \tag{6.2}$$

where $\mu_{\mathbf{w}}$ and $\sigma_{\mathbf{w}}$ are fixed during the encoder training process. Therefore, using equation 6.2, we can explicitly constrain the latent representation output by our encoder to lie in the latent representation space $\mathcal{W}+$ of our decoder G.

Network optimization Our encoder E is trained using two losses. We employ the pixel-wise reconstruction loss $\mathscr{L}_{mse}$ to enforce reconstructed HR images to be close to their HR ground-truth I_{HR},

$$\mathscr{L}_{mse} = \|I_{HR} - G(\mathbf{w}))\|_2. \tag{6.3}$$

In addition, we also introduce the perceptual loss to enforce the feature-wise similarity,

$$\mathscr{L}_{percept} = \|F(I_{HR}) - F(G(\mathbf{w}))\|_2, \tag{6.4}$$

where F denotes the perceptual feature extractor. In our experiments, we extract features from `relu1_1`, `relu2_1`, `relu3_1`, `relu4_1` layers in VGG-19 with equal weights. In our final objective, we also treat the image intensity similarity and feature similarity equally, and the objective is defined as,

$$\mathscr{L}(\theta) = \mathscr{L}_{mse} + \mathscr{L}_{percept}, \tag{6.5}$$

where θ is the trainable parameters of our network. Note that our upsampling decoder and encoder are trained individually and thus our decoder is fixed during training our encoder.

Table 6.1: Comparison with state-of-the-art methods. Results are reported on three benchmarks noted as source → target. 'Source only' denotes the methods only using source dataset for training, while 'one-shot' denotes the methods exploring one-shot exemplar on the target dataset. ↑ indicates that higher is better, and ↓ that lower is better.

	Method	CelebA → ExtYaleB				CelebA → MultiPIE				MultiPIE → ExtYaleB			
		LPIPS ↓	FIQ ↑	PSNR ↑	SSIM ↑	LPIPS ↓	FIQ ↑	PSNR ↑	SSIM ↑	LPIPS ↓	FIQ ↑	PSNR ↑	SSIM ↑
source only	Bicubic	0.52	0.31	19.94	0.46	0.55	0.27	17.11	0.39	0.54	0.31	17.70	0.43
	PUSLE [122]	0.40	0.38	20.18	0.46	0.46	0.36	14.63	0.37	0.42	0.27	17.02	0.46
	MTDN [202]	0.39	0.32	17.74	0.45	**0.38**	0.38	18.00	0.52	0.47	0.20	18.67	0.43
	CPGAN [196]	0.40	0.28	17.03	0.47	0.40	0.31	18.61	0.52	0.45	0.24	18.80	0.44
	DAP-FSR (Ours)	**0.38**	**0.41**	**20.39**	**0.49**	**0.38**	**0.40**	**19.15**	**0.54**	**0.41**	**0.34**	**19.28**	**0.46**
one-shot	PULSE+ASM [1]	0.44	0.32	20.47	0.47	0.49	0.32	17.87	0.41	0.44	0.23	17.89	0.43
	MTDN+ASM	0.42	0.27	19.01	0.48	0.44	0.33	19.38	0.53	0.52	0.25	19.11	0.47
	CPGAN+ASM	0.49	0.26	18.42	0.42	0.49	0.29	19.29	0.55	0.51	0.23	19.19	0.49
	DAP-FSR (Ours)	**0.36**	**0.46**	**22.32**	**0.55**	**0.36**	**0.44**	**21.00**	**0.61**	**0.39**	**0.40**	**20.43**	**0.51**

6.3.2 Peeking at One-Shot Exemplar

Towards target-domain image generation Benefiting from the encoder design in our DAP-FSR network, we can encode the given one-shot target domain HR image I^t_{HR} into a latent representation $\mathbf{w}^t$. However, using only one-shot exemplar does not suffice to transfer our decoder to the target domain, and will lead to an over-fitting problem. As explained in [198], the latent codes of the StyleGAN control the coarse, medium, fine attributes of generated images at different style layers. Thus, we also regard the latent code $\mathbf{w}^t$ as an interpretable representation of a target domain face. Moreover, we can generate a large number of domain-specific (*i.e.*, style-consistent) face images with I^t_{HR}. Specifically, for a latent code $\mathbf{w}^s$ randomly sampled from the latent representation manifold of the source domain, we mix it with $\mathbf{w}^t$ in a layer-wise manner so that a generated image I^m inherits the the target domain style from I^t. The mixing procedure is defined as:

$$\mathbf{w}^m_i = (1-\alpha_i)\mathbf{w}^t_i + \alpha_i \mathbf{w}^s_i, \tag{6.6}$$

where $\alpha \in \mathbb{R}^l$ is a layer-wise soft weight for mixing latent representations. In this manner, we effectively enlarge the number of target domain examples from the given one-shot exemplar by $G(\mathbf{w}^m)$. In Figure 6.3, compared with a style transfer based method (*i.e.*, ASM [1]), our method is able to generate more natural style-consistent images while preserving the identity.

Learning soft mixing weight When mixing the latent representations of random sampled $\mathbf{w}^s$ and the target sample $\mathbf{w}^t$, we preserve the image content information by applying a feature-wise intensity consistent loss $\mathscr{L}_c$ and enforce the domain information

to be transferred by employing a style similarity loss $\mathscr{L}_s$. Here, we learn a soft weight α to mix the latent codes of the source and target domain instead of manually selecting a certain layer, and the optimization process is formulated as,

$$\mathscr{L}(\alpha) = \mathscr{L}_c + \mathscr{L}_s, \tag{6.7}$$

$$\mathscr{L}_c = \|F(G(\mathbf{w}^m)) - F(G(\mathbf{w}^s))\|_2, \tag{6.8}$$

$$\begin{aligned}\mathscr{L}_s = &\|\mu(F(I^t) - \mu(F(G(\mathbf{w}^m)))\|_2 + \\ &\|\sigma(F(I^t)) - \sigma(F(G(\mathbf{w}^m)))\|_2.\end{aligned} \tag{6.9}$$

where μ and σ denote the mean and variance of the extracted features respectively, and F is the same perceptual extractor in Eq. equation 6.4.

Model updating by constrained adaptation After we generate a batch of random images exhibiting the same target domain style, our next step is to adapt our model towards the target domain. The most straightforward way is to fine-tune the entire decoder G directly on our generated target-domain alike samples. However, when the number of training examples is limited, especially in our case, fine-tuning the whole network weights often leads to over-fitting and may potentially destroy the learned knowledge prior in G. Instead of fine-tuning the entire decoder weights, we constrain the fine-tuning on a subset of the decoder parameters. To be specific, we only adapt the affine transform parameters in the AdaIN module. By restricting the trainable parameters, our model can be effectively adapted to the target domain while preserving the semantic knowledge, *i.e.*, natural face structure. The overall pipeline of our algorithm is illustrated in Algorithm 2.

6.3.3 Training and Inference

Our training process consists of two main stages, the procurement stage and development stage. In the procurement stage, we first train our decoder G following the protocols of StyleGAN and then only train the encoder model E on the source dataset by Eq. equation 6.5 while fixing the parameters of G. After training, our DAP-FSR is able to super-resolve HR faces from LR faces with an upscaling factor up to $\times 64$. In the development stage, we peek at the one-shot exemplar from the target domain and adapt our model to the target domain by employing our proposed Algorithm 2. During inference, we test our adapted model on the whole target dataset and report the super-resolution performance. Note that, we only see one-shot image from the target domain and all other testing images are *never seen* during training.

6.4 Experiments

In this section, we conduct extensive experiments to evaluate our DAP-FSR framework. Since we focus on the OSDA-FSR task, we mainly compare with the state-of-the-art in this scenario.

6.4.1 Datasets and Evaluation Protocols

Benchmarks Current FSR benchmarks conduct training and testing within the same domain, and do not support the setting of the cross-domain OSDA-FSR task. Therefore, We propose three benchmarks to evaluate the performance of our DAP-FSR, *i.e.*, CelebA [203] → Multi-PIE [204], CelebA → ExtendedYaleB [205], and Multi-PIE → ExtendedYaleB. In particular, CelebA dataset contains large-scale in-the-wild face images, Multi-PIE and ExtendedYaleB datasets comprise indoor face images captured in different poses and illumination conditions. We select 10 different illumination and pose condition data splits in Multi-PIE and ExtendedYaleB, respectively. The adaptation performance is evaluated with a given exemplar in each split and then the final reported performance is averaged over all the splits.

Evaluation metrics We report the quantitative results using the average Peak Single-to-Noise Ratio (PSNR), Structural SIMilarity scores (SSIM) following the common FSR practice [193, 196]. Furthermore, we also employ the Learned Perceptual Image Patch Similarity (LPIPS) [206] and Face Image Quality (FIQ) [207] to evaluate the quality and authenticity of super-resolved faces. The PSNR, SSIM, LPIPS metrics are calculated between the reconstructed HR images I_{SR} and the ground-truth HR images I_{HR}. The FIQ is a non-reference metric for face quality assessment, which is calculated only on I_{SR}.

6.4.2 Implementation Details

In our experiments, we crop the aligned faces and resize them to 128×128 pixels to achieve ground-truth HR images. In real-world applications, we do not assume that the input LR faces are perfectly aligned. Following [202], we apply affine transformations, including rotations, translations and scaling, to HR faces and then downsample them to 16×16 pixels as our LR face images. We use the author-provided codes of PULSE [122], MTDN [202] and CPGAN [196]. For comparison fairness, we adopt the same training protocols for all the methods. To alleviate the influence of the selected one-shot exemplar,

Figure 6.4: Comparisons with state-of-the-art methods on CelebA→ExtYaleB, CelebA→MultiPIE and MultiPIE→ExtYaleB benchmarks under the OSDA-FSR setting. Our method achieves high-quality, style-consistent HR faces and is also robust against unaligned LR inputs.

we run the proposed method for ten times with different randomly selected one-shot exemplars in each task and report the averaged results.

6.4.3 Comparisons with the State-of-the-Art

Qualitative comparisons We first conduct qualitative comparisons with the state-of-the-art methods on three OSDA-FSR benchmarks in Figure 6.4.

CPGAN [196] and MTDN [202] can super-resolve LR images well and deal with unaligned LR input faces successfully in the source domain. However, these methods do not take the domain gap into account, and lack an efficient mechanism to address

LR images from a new domain. Therefore, their final reconstructed HR images from target domain LR faces suffer from severe artifacts. Although collecting a large number of target domain data and then re-training the networks can solve the above issue, doing so is time-consuming and does not provide a data-efficient solution to OSDA-FSR.

PULSE [122] traverses the high-resolution face image manifold and searches images whose downsampled versions are close to the given LR images. Although realistic images are achieved, this method requires input LR images to be perfectly pre-aligned. When LR images are unaligned, the reconstructed HR images are enforced to match the intensities of LR faces. This will lead to severe changes of face identities, as seen in Figure 6.4. Moreover, PULSE does not consider the domain gap. Due to the data distribution shift between the source and target domains, PULSE fails to super-resolve HR faces sharing the same style as the target domain images.

In contrast, as seen in Figure 6.4, our method achieves superior performance compared to the other competing methods. Although input LR images are unaligned, our DAP-FSR still produces visually appealing HR faces which are close to their HR ground-truth. Notably, our upsampled faces also exhibit style-consistency with respect to the given one-shot target domain exemplar. This demonstrates the transfer ability of our method. Note that our method is actually able to super-resolve LR faces with an upscaling factor up to 64×, and for fair comparisons with the state-of-the-art methods, we only show HR faces in the same resolution as other methods. To the best of our knowledge, our DAP-FSR network is *the first attempt to super-resolve cross-domain LR images with only one target-domain exemplar*, and achieves superior super-resolution results.

To further validate the generalization ability, in Figure 6.5, we show the FSR results of tiny faces *in-the-wild* [2] under *real-world unconstrained conditions*, where the ground-truth HRs are unavailable. Here, LR faces may undergo different poses, blurs, noises, etc. All the models are trained on the CelebA source dataset and adapted to the target domain using the given one-shot HR example. Moreover, in Figure 6.6, we also conduct cross-domain FSR experiments on near infrared (NIR) face images [3] as a target domain. Our DAP-FSR still outperforms the other competing methods, demonstrating the generalization ability of our method.

Quantitative comparisons As indicated in Table 6.1, we report the LPIPS, FIQ, PSNR and SSIM metrics on three OSDA-FSR benchmarks, respectively. Our proposed DAP-FSR outperforms the state-of-the-art methods significantly, especially on the perceptually-driven metrics, *i.e.*, LPIPS and FIQ. This indicates that our super-resolved target domain HR faces not only resemble their ground-truth but also are photo-realistic. More impor-

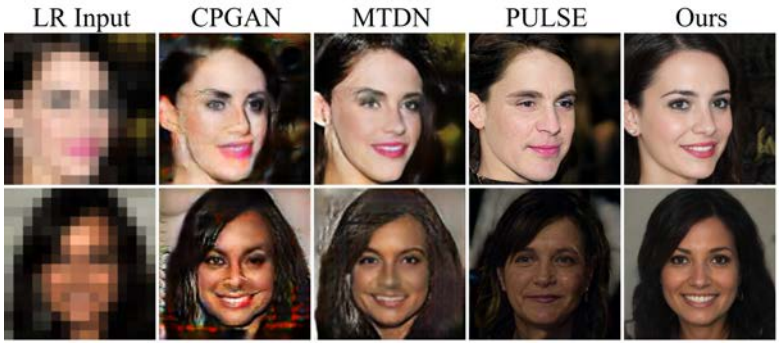


Figure 6.5: Comparisons with state-of-the-art methods on tiny faces in-the-wild [2] under real-world unconstrained conditions.

tantly, our DAP-FSR consistently performs better than other methods on all the benchmarks. Thanks to our dedicated network design, we are able to align and upsample target domain LR faces, simultaneously. In particular, DAP-FSR reconstructs high-quality face images and outperforms the second best method PULSE on unaligned images by a margin of +43% (0.32 → 0.46) in FIQ on the benchmark CelebA→ExtendedYaleB.

To address the domain gap, a straightforward idea is fine-tuning the source-trained FSR model with the augmented target samples. Thus, we employ a style-transfer-based method ASM [1] to augment new training samples from the one-shot target domain exemplar, and then fine-tune the FSR models. We name these the combination as +ASM in Table 6.1. As indicated by Table 6.1, applying style transfer cannot fully establish the facial detail correspondences between the source and target domains, thus leading to performance degradation.

Furthermore, benefiting from our designed one-shot adaptation algorithm, we transfer our network to the target domain effectively. Therefore, our quantitative results are better than the results of MTDN+ASM and PULSE+ASM. Owing to our encoder-decoder design, our method is also more efficient and effective compared to the decoder-only based method PULSE. After training, our DAP-FSR hallucinates LR faces in a feed-forward manner and runs ×150 faster than PULSE, which provides a high application potential in the real-world scenario.

6.4.4 Ablation Analysis

In our ablation analysis, we conduct all the experiments on the CelebA→ExtendedYaleB benchmark.

Effectiveness of network design We analyze the effect of each component in our

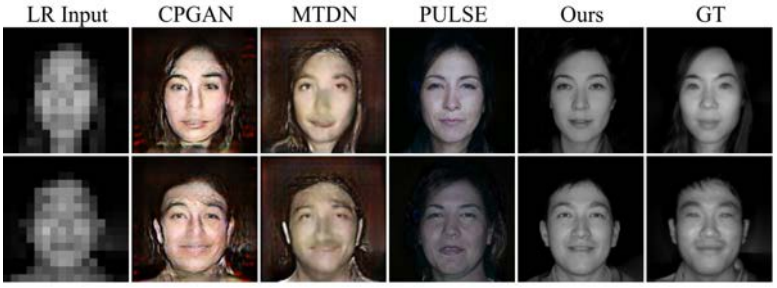


Figure 6.6: Comparisons with state-of-the-art methods on near-infrared (NIR) sensor captured faces [3].

network design in Table 6.2. Compared to a straightforward approach that predicts the latent representations at the end of the backbone, our network adaptively explores the abundant multi-scale features (Config A). It is a long-standing shortcoming that CNN is sensitive to rotations. Our multiple ISTN design effectively handle this problem (Config B), thus being robust against unaligned LR images. We also illustrate that it is vital to explicitly constrain the predicted latent representations on the manifold (Config C).

Effectiveness of one-shot domain adaptation Table 6.2 indicates the impact of each component in Algorithm 2 on the OSDA-FSR performance. In our method, we effectively enrich the training samples by mixing the latent representations between the source and target domain faces (Config D). Compared to the configuration without exploring the one-shot exemplar (Config C), we observe that Config D achieves better super-resolution performance. This implies our method fully exploits the one-shot target exemplar to bridge the domain gap.

By applying the soft mixing weight (Config E), we further improve the super-resolution performance. This indicates that our soft mixing strategy is more effective than simply replacing the last three final layers of the latent representations between the source and target domain images as done in Config D. As fine-tuning the whole decoder network may lead to over-fitting and destroy the learned face priors, we constrain the optimization space and only modify the AdaIN parameters to improve performance (Config F).

We also compare with other target domain augmentation methods, including Style Transfer and ASM. Specifically, these are employed to enlarge the target domain examples and then we constrained fine-tune our model using the augmented data. As indicated in Table 6.3, our method significantly facilitates the model adapting to the target domain, thus achieve better super-resolution performance.

Table 6.2: Ablations on different configurations of the network architecture (A,B,C) and different configurations of the adaptation algorithm (D,E,F). ↑ indicates the higher the better, and ↓ indicates the lower the better.

Configuration	**CelebA → ExtendedYaleB**			
	LPIPS ↓	FIQ ↑	PSNR ↑	SSIM ↑
Baseline network	0.48	0.28	17.64	0.44
A + Multi-scale features	0.46	0.30	17.82	0.44
B + Multi-STN modules	0.43	0.34	18.41	0.45
C + Predict offset scale	0.38	0.41	20.39	0.49
D + Style mixing examples	0.38	0.41	21.97	0.52
E + Soft mixing weight	0.38	0.42	22.10	0.54
F + Constrained adaptation	0.36	0.46	22.32	0.55

Table 6.3: Comparisons on one-shot adaptation augmentation strategies. ↑ indicates the higher the better, and ↓ the lower the better.

Methods	**CelebA → ExtendedYaleB**			
	LPIPS ↓	FIQ ↑	PSNR ↑	SSIM ↑
Direct fine-tuning	0.44	0.30	20.11	0.45
Style Transfer [135]	0.42	0.37	20.16	0.46
ASM [1]	0.40	0.38	20.71	0.50
DAP-FSR (Ours)	0.36	0.46	22.32	0.55

6.5 Summary

In this section, we addressed a more challenging and practical face super-resolution task, where a domain gap between the training and testing data exists. To tackle this problem, we proposed a new Domain-Aware Pyramid-based Face Super-Resolution network (DAP-FSR) that is able to super-resolve unaligned low-resolution ones from a target domain effectively by leveraging only one target domain exemplar. Our approach bridges the domain gap by fully exploiting the given exemplar from the target domain as well as our designed soft mixing strategy which significantly enlarges the number of the training samples. Extensive experiments demonstrate our method is able to super-resolve cross-domain LR faces and outperforms the state-of-the-art methods significantly. We hope that our work will also motivate future research on the low-shot FSR task.

CHAPTER 7

CONCLUSION

In this thesis, we propose data-efficient algorithms to alleviate the data requirements for deep learning networks in terms of data quantity and quality. More specifically, we present data-efficient algorithms for visual understanding tasks that exploit (1) learning with noise, (2) few-shot learning and zero-shot learning, and (3) transferring prior knowledge. Our major contributions can be summarized as followings,

Human Parsing from Learning with Noise Perspective We propose to tackle the challenging human parsing task by considering the label noises existing in the ground-truth masks. To the best of our knowledge, this is a new perspective in this research area, which is not well explored before. We propose a simple yet effective noise-tolerant approach named SCHP for alleviating the existing label noises accordingly. By alternatively performing model aggregating and label refining in an online manner, SCHP could mutually promote the model performance and label accuracy. Our SCHP is model-agnostic and thus can be applied to various human parsing frameworks. Extensive ablation experiments demonstrate the generalization ability and the superiority of the proposed SCHP. Benefiting from the proposed SCHP, this work achieved a new state-of-the-art performance on six single/multiple human parsing benchmarks and won the winner prize of all three human parsing tracks in the 3rd Look Into Person Challenge.

Generalized Few-shot Scene Parsing from Meta-Learning Perspective We advance the few-shot segmentation paradigm towards a more challenging yet general scenario, *i.e.*, generalized few-shot scene parsing. We present a generic framework named Meta Parsing Networks (MPNet). Our MPNet is a generic framework for performing

the generalized few-shot scene parsing task. We conduct experiments on two newly constructed generalized few-shot scene parsing benchmarks, called *GFSP-Cityscapes* and *GFSP-Pascal-Context*. Extensive ablation studies and comparisons well demonstrate the effectiveness and generalization ability of our proposed MPNet.

Zero-shot Semantic Segmentation by Cross-modal Knowledge Transferring We propose Consistent Structural Relation Learning (CSRL) framework to tackle the challenging generalized zero-shot semantic segmentation task. We propose a semantic-visual structural generator by integrating both feature generating and relation learning in a unified network architecture. We conduct extensive experiments on two GZS3 benchmarks based on Pascal-VOC and Pascal-Context datasets. The proposed CSRL outperforms existing state-of-the-art methods by a large margin, resulting in ~7-12% on Pascal-VOC and ~2-5% on Pascal-Context.

Cross-domain Face Super-resolution via One-shot Exemplar We propose a novel domain-aware pyramid-based face super-resolution network, named DAP-FSR network, to efficiently upsample cross-domain LR face images by peeking at one-shot target domain example. We present a simple yet effective domain-aware latent mixing and model adaptation algorithm (DAMMA) to adapt our DAP-FSR to the target domain. Our DAMMA generates target-style alike faces to adapt the upsampling decoder in DAP-FSR by fully exploiting the one-shot example. To the best of our knowledge, our method is the first attempt to super-resolve cross-domain LR face images, making our method more practical. Our proposed DAP-FSR can effectively adapt to a target domain and is also robust to unaligned LR faces. Experiments on three constructed cross-domain face super-resolution benchmarks validate the superior performance of our proposed approach compared to the state-of-the-art methods.

In summary, standing on the shoulders of former practices, we have taken multiple stages in this thesis to investigate deep learning approaches for data-efficient visual understanding. The successful application of data-efficient machine learning techniques into computer vision tasks will make the deep learning models perform well and generalize with limited annotated data examples. Although the works introduced in this thesis achieve an initial success for the data-efficient visual understanding tasks, there is still a long way. We can observe that there is still a significant performance gap between the results of the fully-supervised paradigm and the data-efficient paradigm. Thus, more effective data-efficient learning-based algorithms are still required to alleviate this gap. We hope that our efforts will motivate more researchers and ease future research on data-efficient visual understanding.

BIBLIOGRAPHY

[1] Y. Luo, P. Liu, T. Guan, J. Yu, and Y. Yang, "Adversarial style mining for one-shot unsupervised domain adaptation," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.

[2] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "Finding tiny faces in the wild with generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 21–30.

[3] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The casia nir-vis 2.0 face database," in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 348–353.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 630–645.

[5] ——, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[6] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 2018.

[7] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881–2890.

[8] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth

16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[10] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning (ICML)*, 2017, pp. 1126–1135.

[11] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2. Lille, 2015.

[12] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2016, pp. 3630–3638.

[13] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4077–4087.

[14] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1199–1208.

[15] L. Zhang, T. Xiang, and S. Gong, "Learning a deep embedding model for zero-shot learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2021–2030.

[16] H. Jiang, R. Wang, S. Shan, and X. Chen, "Transferable contrastive network for generalized zero-shot learning," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 9765–9774.

[17] D. Mandal, S. Narayan, S. K. Dwivedi, V. Gupta, S. Ahmed, F. S. Khan, and L. Shao, "Out-of-distribution detection for generalized zero-shot action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9985–9993.

[18] J. Gao, T. Zhang, and C. Xu, "I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs," in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 33, 2019, pp. 8303–8311.

[19] S. Rahman, S. Khan, and F. Porikli, "Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts," in *Asian Conference on Computer Vision (ACCV)*, 2018, pp. 547–563.

[20] A. Bansal, K. Sikka, G. Sharma, R. Chellappa, and A. Divakaran, "Zero-shot object detection," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 384–400.

[21] Y. Xian, S. Choudhury, Y. He, B. Schiele, and Z. Akata, "Semantic projection network for zero-and few-label semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8256–8265.

[22] M. Bucher, T.-H. Vu, M. Cord, and P. Pérez, "Zero-shot semantic segmentation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[23] Y. Li, D. Wang, H. Hu, Y. Lin, and Y. Zhuang, "Zero-shot recognition using dual visual-semantic mapping paths," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3279–3287.

[24] S. Changpinyo, W.-L. Chao, and F. Sha, "Predicting visual exemplars of unseen classes for zero-shot learning," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3476–3485.

[25] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, "Synthesized classifiers for zero-shot learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5327–5336.

[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 2672–2680.

[27] X. Wang and A. Gupta, "Generative image modeling using style and structure adversarial networks," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 318–335.

[28] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5542–5551.

[29] V. Kumar Verma, G. Arora, A. Mishra, and P. Rai, "Generalized zero-shot learning via synthesized examples," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4281–4289.

[30] B. Frénay and M. Verleysen, "Classification in the presence of label noise: a survey," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 5, pp. 845–869, 2013.

[31] J. Zhang, X. Wu, and V. S. Sheng, "Learning from crowdsourced labeled data: a survey," *Artificial Intelligence Review*, vol. 46, no. 4, pp. 543–576, 2016.

[32] X. Chen and A. Gupta, "Webly supervised learning of convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1431–1439.

[33] L. Cheng, X. Zhou, L. Zhao, D. Li, H. Shang, Y. Zheng, P. Pan, and Y. Xu, "Weakly supervised learning with side information for noisy labeled images," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 306–321.

[34] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, "Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *International Conference on Machine Learning (ICML)*. PMLR, 2018, pp. 2304–2313.

[35] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Robust learning by self-transition for handling noisy labels," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD)*, 2021, pp. 1490–1500.

[36] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 1195–1204.

[37] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima and better generalization," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

[38] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on Challenges in Representation Learning, ICML*, vol. 3, 2013, p. 2.

[39] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich, "Training deep neural networks on noisy labels with bootstrapping," *arXiv preprint arXiv:1412.6596*, 2014.

[40] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 4694–4703.

[41] M.-Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen, and J. Kautz, "Few-shot unsupervised image-to-image translation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 10 551–10 560.

[42] S. Motiian, Q. Jones, S. Iranmanesh, and G. Doretto, "Few-shot adversarial domain adaptation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 6670–6680.

[43] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017, pp. 3722–3731.

[44] S. Benaim and L. Wolf, "One-shot unsupervised cross domain translation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 2104–2114.

[45] S. Wang, L. Li, Y. Ding, C. Fan, and X. Yu, "Audio2head: Audio-driven one-shot talking-head generation with natural head motion," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

[46] Y. Liu, J. Lee, M. Park, S. Kim, E. Yang, S. J. Hwang, and Y. Yang, "Learning to propagate labels: Transductive propagation network for few-shot learning," in *ICLR*, 2019.

[47] Q. Feng, G. Kang, H. Fan, and Y. Yang, "Attract or distract: Exploit the margin of open set," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 7990–7999.

[48] Q. Feng, Z. Yang, P. Li, Y. Wei, and Y. Yang, "Dual embedding learning for video instance segmentation," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019.

[49] Q. Feng, Y. Wu, H. Fan, C. Yan, M. Xu, and Y. Yang, "Cascaded revision network for novel object captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[50] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, vol. 28, no. 4, pp. 594–611, 2006.

[51] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *arXiv preprint arXiv:1803.02999*, 2018.

[52] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," 2016.

[53] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1125–1134.

[54] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2223–2232.

[55] Y. Yuan and J. Wang, "Ocnet: Object context network for scene parsing," *arXiv preprint arXiv:1809.00916*, 2018.

[56] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *European Conference on Computer Vision (ECCV)*, 2020.

[57] L. Huang, Y. Yuan, J. Guo, C. Zhang, X. Chen, and J. Wang, "Interlaced sparse self-attention for semantic segmentation," *arXiv preprint arXiv:1907.12273*, 2019.

[58] P. Li, Y. Xu, Y. Wei, and Y. Yang, "Self-correction for human parsing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[59] P. Li, X. Dong, X. Yu, and Y. Yang, "When humans meet machines: Towards efficient segmentation networks," *Proceedings of the British Machine Vision Conference (BMVC)*, 2020.

[60] G. Li, G. Kang, W. Liu, Y. Wei, and Y. Yang, "Content-consistent matching for domain adaptive semantic segmentation," in *European Conference on Computer Vision (ECCV)*, 2020.

[61] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2507–2516.

[62] Y. Luo, P. Liu, T. Guan, J. Yu, and Y. Yang, "Adversarial style mining for one-shot unsupervised domain adaptation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[63] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, and S. Yan, "Stc: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE Transactions on Pattern Recognition and Machine Intelligence*, vol. 39, no. 11, pp. 2314–2320, 2016.

[64] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1568–1576.

[65] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7268–7277.

[66] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "Panet: Few-shot image semantic segmentation with prototype alignment," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 9197–9206.

[67] A. Li, W. Huang, X. Lan, J. Feng, Z. Li, and L. Wang, "Boosting few-shot learning with adaptive margin loss," *arXiv preprint arXiv:2005.13826*, 2020.

[68] P. Li, Y. Wei, and Y. Yang, "Meta parsing networks: Towards generalized few-shot scene parsing with adaptive metric learning," in *ACM International Conference on Multimedia (MM)*, 2020, pp. 64–72.

[69] ——, "Consistent structural relation learning for zero-shot segmentation," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.

[70] P. Li, X. Yu, and Y. Yang, "Super-resolving cross-domain face miniatures by peeking at one-shot exemplar," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 4469–4479.

[71] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.

[72] B. Cheng, L.-C. Chen, Y. Wei, Y. Zhu, Z. Huang, J. Xiong, T. S. Huang, W.-M. Hwu, and H. Shi, "Spgnet: Semantic prediction guidance for scene parsing," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 5218–5228.

[73] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 603–612.

[74] J. Jiao, Y. Wei, Z. Jie, H. Shi, R. W. Lau, and T. S. Huang, "Geometry-aware distillation for indoor semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2869–2878.

[75] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2961–2969.

[76] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal on Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[77] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 740–755.

[78] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 633–641.

[79] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene un-

derstanding," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3213–3223.

[80] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision (ECCV)*, 2012, pp. 746–760.

[81] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 567–576.

[82] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[83] Z. Yang, P. Li, Q. Feng, Y. Wei, and Y. Yang, "Going deeper into embedding learning for video object segmentation," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[84] X. Pan, P. Li, Z. Yang, H. Zhou, C. Zhou, H. Yang, J. Zhou, and Y. Yang, "In-n-out generative learning for dense unsupervised video segmentation," *arXiv preprint arXiv:2203.15312*, 2022.

[85] Q. Feng, Y. Wei, M. Cheng, and Y. Yang, "Decoupled spatial temporal graphs for generic visual grounding," *arXiv preprint arXiv:2103.10191*, 2021.

[86] Z. Zheng, Y. Wei, and Y. Yang, "University-1652: A multi-view multi-source benchmark for drone-based geo-localization," in *Proceedings of the 28th ACM international conference on Multimedia (MM)*, 2020.

[87] P. Li, P. Pan, P. Liu, M. Xu, and Y. Yang, "Hierarchical temporal modeling with mutual distance matching for video based person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[88] P. Li, Y. Xu, Y. Wei, and Y. Yang, "Self-correction for human parsing," *arXiv preprint arXiv:1910.09777*, 2019.

[89] X. Liang, X. Shen, D. Xiang, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with local-global long short-term memory," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3185–3193.

[90] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 932–940.

[91] X. Liang, K. Gong, X. Shen, and L. Lin, "Look into person: Joint body parsing & pose estimation network and a new benchmark," *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, vol. 41, no. 4, pp. 871–885, 2018.

[92] F. Xia, P. Wang, X. Chen, and A. L. Yuille, "Joint multi-person pose estimation and semantic part segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6769–6778.

[93] X. Nie, J. Feng, and S. Yan, "Mutual learning to adapt for joint human parsing and pose estimation," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 502–517.

[94] T. Ruan, T. Liu, Z. Huang, Y. Wei, S. Wei, and Y. Zhao, "Devil in the details: Towards accurate single and multiple human parsing," in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 33, 2019, pp. 4814–4821.

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

[96] J. Zhao, J. Li, Y. Cheng, T. Sim, S. Yan, and J. Feng, "Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing," in *ACM International Conference on Multimedia (MM)*, 2018, pp. 792–800.

[97] J. Zhao, J. Li, H. Liu, S. Yan, and J. Feng, "Fine-grained multi-human parsing," *International Journal on Computer Vision (IJCV)*, 2019.

[98] K. Gong, Y. Gao, X. Liang, X. Shen, M. Wang, and L. Lin, "Graphonomy: Universal human parsing via graph transfer learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[99] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei, "Flow-guided feature aggregation for video object detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 408–417.

[100] Q. Zhou, X. Liang, K. Gong, and L. Lin, "Adaptive temporal encoding network for video instance-level human parsing," in *ACM International Conference on Multimedia (MM)*, 2018.

[101] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," *arXiv preprint arXiv:1709.03410*, 2017.

[102] T. Hu, P. Yang, C. Zhang, G. Yu, Y. Mu, and C. G. Snoek, "Attention-based multi-context guiding for few-shot semantic segmentation," in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 33, 2019, pp. 8441–8448.

[103] P. Tian, Z. Wu, L. Qi, L. Wang, Y. Shi, and Y. Gao, "Differentiable meta-learning model for few-shot semantic segmentation," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[104] C. Zhang, G. Lin, F. Liu, J. Guo, Q. Wu, and R. Yao, "Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 9587–9595.

[105] M. Siam, B. N. Oreshkin, and M. Jagersand, "Amp: Adaptive masked proxies for few-shot segmentation," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 5249–5258.

[106] C. Zhang, G. Lin, F. Liu, R. Yao, and C. Shen, "Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5217–5226.

[107] K. Nguyen and S. Todorovic, "Feature weighting and boosting for few-shot segmentation," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 622–631.

[108] Q. Feng, "Connecting perception with cognition for deep representations learning," Ph.D. dissertation, 2021.

[109] X. Wang and X. Tang, "Hallucinating face by eigentransformation," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 35, no. 3, pp. 425–434, 2005.

[110] C. Liu, H.-Y. Shum, and W. T. Freeman, "Face hallucination: Theory and practice," *International Journal of Computer Vision (IJCV)*, vol. 75, no. 1, pp. 115–134, 2007.

[111] S. Kolouri and G. K. Rohde, "Transport-based single frame super resolution of very low resolution face images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4876–4884.

[112] X. Ma, J. Zhang, and C. Qi, "Hallucinating face by position-patch," *Pattern Recognition*, vol. 43, no. 6, pp. 2224–2236, 2010.

[113] M. F. Tappen and C. Liu, "A bayesian approach to alignment-based image hallucination," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2012, pp. 236–249.

[114] C.-Y. Yang, S. Liu, and M.-H. Yang, "Hallucinating compressed face images," *International Journal of Computer Vision*, vol. 126, no. 6, pp. 597–614, 2018.

[115] S. Zhu, S. Liu, C. C. Loy, and X. Tang, "Deep cascaded bi-network for face hallucination," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2016, pp. 614–630.

[116] X. Yu and F. Porikli, "Ultra-resolving face images by discriminative generative networks," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2016, pp. 318–333.

[117] ——, "Face hallucination with tiny unaligned images by transformative discriminative neural networks," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 4327–4333.

[118] X. Yu, B. Fernando, R. Hartley, and F. Porikli, "Semantic face hallucination: Super-resolving very low-resolution face images with supplementary attributes," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI*, 2019.

[119] Y. Zhang, I. Tsang, Y. Luo, C. Hu, X. Lu, and X. Yu, "Recursive copy and paste gan: Face hallucination from shaded thumbnails," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[120] H. Huang, R. He, Z. Sun, and T. Tan, "Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2017, pp. 1689–1697.

[121] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "Fsrnet: End-to-end learning face super-resolution with facial priors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2492–2501.

[122] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "Pulse: Self-supervised photo upsampling via latent space exploration of generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2437–2445.

[123] Y. Zhang, I. W. Tsang, J. Li, P. Liu, X. Lu, and X. Yu, "Face hallucination with finishing touches," *IEEE Transactions on Image Processing (TIP)*, vol. 30, pp. 1728–1743, 2021.

[124] X. Yu, B. Fernando, R. Hartley, and F. Porikli, "Super-resolving very low-resolution face images with supplementary attributes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 908–917.

[125] X. Yu, B. Fernando, B. Ghanem, F. Porikli, and R. Hartley, "Face super-resolution guided by facial component heatmaps," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2018, pp. 217–233.

[126] A. Bulat, J. Yang, and G. Tzimiropoulos, "To learn image super-resolution, use a gan to learn how to do image degradation first," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2018, pp. 185–200.

[127] A. Bulat and G. Tzimiropoulos, "Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 109–117.

[128] K. Zhang, Z. Zhang, C.-W. Cheng, W. H. Hsu, Y. Qiao, W. Liu, and T. Zhang, "Super-identity convolutional neural network for face hallucination," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 183–198.

[129] F. Shiri, X. Yu, F. Porikli, R. Hartley, and P. Koniusz, "Identity-preserving face recovery from stylized portraits," *International Journal of Computer Vision*, vol. 127, no. 6-7, pp. 863–883, 2019.

[130] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[131] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.

[132] Q. Feng, Y. Luo, K. Luo, and Y. Yang, "Look, evolve and mold: Learning 3d shape manifold via single-view synthetic data," *arXiv e-prints*, pp. arXiv–2103, 2021.

[133] L. Yu, Q. Feng, Y. Qian, W. Liu, and A. G. Hauptmann, "Zero-virus: Zero-shot vehicle route understanding system for intelligent transportation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 594–595.

[134] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 35–51.

[135] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1501–1510.

[136] C. Yang and S.-N. Lim, "One-shot domain adaptation for face generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5921–5930.

[137] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, vol. 39, no. 4, pp. 640–651, April 2017.

[138] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, vol. 40, no. 4, pp. 834–848, 2017.

[139] X. Liang, C. Xu, X. Shen, J. Yang, S. Liu, J. Tang, L. Lin, and S. Yan, "Human parsing with contextualized convolutional neural network," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1386–1394.

[140] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with graph lstm," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 125–143.

[141] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin, "Instance-level human parsing via part grouping network," in *European Conference on Computer Vision (ECCV)*, September 2018, pp. 770–785.

[142] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille, "Detect what you can: Detecting and representing objects using holistic models and body parts," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1971–1978.

[143] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[144] L. N. Smith, "Cyclical learning rates for training neural networks," in *WACV*, 2017, pp. 464–472.

[145] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," in *International Conference on Learning Representations (ICLR)*, 2017.

[146] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *International Conference on Image Processing (ICIP)*, 2017, pp. 3645–3649.

[147] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3640–3649.

[148] Y. Luo, Z. Zheng, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Macro-micro adversarial network for human parsing," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 418–434.

[149] X. Liu, M. Zhang, W. Liu, J. Song, and T. Mei, "Braidnet: Braiding semantics and details for accurate human parsing," in *ACM International Conference on Multimedia (MM)*, 2019, pp. 338–346.

[150] W. Wang, Z. Zhang, S. Qi, J. Shen, Y. Pang, and L. Shao, "Learning compositional neural information fusion for human parsing," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 5703–5713.

[151] F. Xia, P. Wang, L.-C. Chen, and A. L. Yuille, "Zoom better to see clearer: Human part segmentation with auto zoom net," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 648–663.

[152] B. Zhu, Y. Chen, M. Tang, and J. Wang, "Progressive cognitive human parsing," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[153] H.-S. Fang, G. Lu, X. Fang, J. Xie, Y.-W. Tai, and C. Lu, "Weakly and semi supervised human body part parsing via pose-guided knowledge transfer," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[154] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan, "Deep human parsing with active template regression," *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, vol. 37, no. 12, pp. 2402–2414, 2015.

[155] X. Luo, Z. Su, J. Guo, G. Zhang, and X. He, "Trusted guidance pyramid network for human parsing," in *ACM International Conference on Multimedia (MM)*, 2018, pp. 654–662.

[156] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[157] M. Berman, A. Rannen Triki, and M. B. Blaschko, "The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4413–4421.

[158] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520.

[159] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, "High-resolution representations for labeling pixels and regions," *arXiv preprint arXiv:1904.04514*, 2019.

[160] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 102–118.

[161] Q. Li, A. Arnab, and P. H. Torr, "Holistic, instance-level human parsing," in *BMVC*, 2017.

[162] F. Massa and R. Girshick, "maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch," https://github.com/facebookresearch/maskrcnn-benchmark, 2018.

[163] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2349–2358.

[164] L. Yang, Q. Song, Z. Wang, and M. Jiang, "Parsing r-cnn for instance-level human analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 364–373.

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

[166] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, "A kernel method for the two-sample-problem," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2007, pp. 513–520.

[167] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.

[168] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7794–7803.

[169] N. Dong and E. Xing, "Few-shot semantic segmentation with prototype learning." in *British Machine Vision Conference (BMVC)*, vol. 3, no. 4, 2018.

[170] K. Rakelly, E. Shelhamer, T. Darrell, A. Efros, and S. Levine, "Conditional networks for few-shot semantic segmentation," 2018.

[171] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[172] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1520–1528.

[173] X. Liang, Z. Hu, H. Zhang, L. Lin, and E. P. Xing, "Symbolic graph reasoning meets convolutions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 1853–1863.

[174] S. J. Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, and B. Schiele, "Exploiting saliency for object segmentation from image level labels," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5038–5047.

[175] L. Jing, Y. Chen, and Y. Tian, "Coarse-to-fine semantic segmentation from image-level labels," *IEEE Transactions on Image Processing*, vol. 29, pp. 225–236, 2019.

[176] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1635–1643.

[177] C.-C. Hsu, K.-J. Hsu, C.-C. Tsai, Y.-Y. Lin, and Y.-Y. Chuang, "Weakly supervised instance segmentation using the bounding box tightness prior," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 6582–6593.

[178] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3159–3167.

[179] Z. Shu, X. Shen, S. Xin, Q. Chang, J. Feng, L. Kavan, and L. Liu, "Scribble based 3d shape segmentation via weakly-supervised learning," *IEEE transactions on visualization and computer graphics*, 2019.

[180] H. Xiao, B. Kang, Y. Liu, M. Zhang, and J. Feng, "Online meta adaptation for fast video object segmentation," *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 2019.

[181] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International Conference on Machine Learning (ICML)*, 2014, pp. 1188–1196.

[182] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014, pp. 1532–1543.

[183] O. Russakovsky and L. Fei-Fei, "Attribute learning in large-scale datasets," in *European Conference on Computer Vision (ECCV)*, 2010, pp. 1–14.

[184] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: from pairwise approach to listwise approach," in *International Conference on Machine Learning (ICML)*, 2007, pp. 129–136.

[185] F. Xia, T.-Y. Liu, J. Wang, W. Zhang, and H. Li, "Listwise approach to learning to rank: theory and algorithm," in *International Conference on Machine Learning (ICML)*, 2008, pp. 1192–1199.

[186] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning‚Äîa comprehensive evaluation of the good, the bad and the ugly," *IEEE Transactions on Pattern Recognition and Machine Intelligence*, vol. 41, no. 9, pp. 2251–2265, 2018.

[187] Y. Li, K. Swersky, and R. Zemel, "Generative moment matching networks," in *International Conference on Machine Learning (ICML)*, 2015, pp. 1718–1727.

[188] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2013, pp. 2121–2129.

[189] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 212–220.

[190] Y. Jo and J. Park, "Sc-fegan: Face editing generative adversarial network with user's sketch and color," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[191] W. Zhibo, Y. Xin, L. Ming, W. Quan, Q. Chen, and X. Feng, "Single image based portrait relighting via explicit multiple channel modeling," in *Siggraph Asia*, 2020, pp. 1–13.

[192] L. Li, S. Wang, Z. Zhang, Y. Ding, Y. Zheng, X. Yu, and C. Fan, "Write-a-speaker: Text-based emotional and rhythmic talking-head generation," in *Proceedings*

of the AAAI Conference on Artificial Intelligence (AAAI), vol. 35, no. 3, 2021, pp. 1911–1920.

[193] X. Yu and F. Porikli, "Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3760–3768.

[194] ——, "Imagining the unimaginable faces by deconvolutional networks," *IEEE Transactions on Image Processing (TIP)*, vol. 27, no. 6, pp. 2747–2761, 2018.

[195] X. Yu, F. Shiri, B. Ghanem, and F. Porikli, "Can we see more? joint frontalization and hallucination of unaligned tiny faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 42, no. 9, pp. 2148–2164, 2019.

[196] Y. Zhang, I. W. Tsang, Y. Luo, C.-H. Hu, X. Lu, and X. Yu, "Copy and paste gan: Face hallucination from shaded thumbnails," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 7355–7364.

[197] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in Neural Information Processing Systems(NIPS)*, 2015, pp. 2017–2025.

[198] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2019, pp. 4401–4410.

[199] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8110–8119.

[200] D. Y. Park and K. H. Lee, "Arbitrary style transfer with style-attentional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5880–5888.

[201] J. Zhu, Y. Shen, D. Zhao, and B. Zhou, "In-domain gan inversion for real image editing," in *European Conference on Computer Vision (ECCV)*, 2020.

[202] X. Yu, F. Porikli, B. Fernando, and R. Hartley, "Hallucinating unaligned face images by multiscale transformative discriminative networks," *International Journal of Computer Vision (IJCV)*, vol. 128, no. 2, pp. 500–526, 2020.

[203] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015, p. 3730,Äì3738.

[204] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," in *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, 2008, pp. 1–8.

[205] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 23, no. 6, pp. 643–660, 2001.

[206] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[207] J. Hernandez-Ortega, J. Galbally, J. Fierrez, R. Haraksim, and L. Beslay, "Faceqnet: quality assessment for face recognition based on deep learning," in *2019 International Conference on Biometrics (ICB)*. IEEE, 2019, pp. 1–8.