Part-Aware and Robust Deep Learning Models for Vision Applications

Jiaxu Miao

Supervisor: Prof. Yi Yang

Faculty of Engineering and Information Technology

University of Technology Sydney

Thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

June 2021

Declaration

I, Jiaxu Miao 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.

Jiaxu Miao
June 2021

Acknowledgements

First, I would like to thank my supervisor, Professor Yi Yang. I could not start my doctoral career without his support and help, and it is luckiest for me to pursue a PhD degree under his supervision. I am extremely grateful for his patient guidance, encouragement and selfless support during my doctoral life. He taught me how to start scientific researches and provided helpful advice about my research career. Whenever I met difficulties about academic or personal life, I can get instant help from him.

I would like to thank Dr. Yunchao Wei. He helped me a lot about my research and shared many valuable ideas with me. He also helped me to revise and promote the writings of my papers.

I would like to thank my friend Yu Wu for his help and kindness from the undergraduate period to the doctoral period. I would also like to thank my group members and colleagues, Linchao Zhu, Xin Yu, Pingbo Pan, Ping Liu, Xiaohan Wang, Peike Li, Qianyu Feng, Yutian Lin, Liang Zheng, Xiaojun Chang, Fan Ma, Zhedong Zheng, Hehe Fan, Xuanyi Dong, Yanbin Liu, Hu Zhang, Guang Li, Zhun Zhong, Yawei Luo, Qingji Guan, Guangrui Li, Qi Rao, Ruijie Quan, Tianqi Tang, Yang He, Zongxin Yang, Chen Liang, Xiaolin Zhang, Yuhang Ding, Yunqiu Xu, Youjiang Xu, and many others. Discussions about the research topics with them help me a lot.

I would like to thank my parents, Shiying Miao and Xiufen Wang for their selfless love. They gave me any support when I needed it. They gave me strength and courage when I faced difficulties and challenges in my life.

I would like to thank my beloved wife, Danwen Sun for her sweet love. She is my best friend, my soulmate, and my heart. She always encourages me with her mild sound whenever I feel upset. She taught me to keep an optimistic attitude to life. Thanks for her patience and tremendous help.

Abstract

With the development of deep learning, the deep models based on neural networks play an important role in vision applications. This dissertation focuses on two limitations of previous deep models. First, early approaches for vision tasks usually focus on global representations, while ignoring the discriminative partial features. However, partial representations provide sufficient recognition information for vision tasks and need to be well developed. Second, deep learning models are eager for massive data with labels, which is hard to acquire. The lack of labeled data inherently introduces uncertainty in deep models. Thus, a robust model should not only provide accurate predictions but also estimate uncertainty precisely.

This dissertation presents part-aware and robust deep models for some important vision applications, *i.e.*, the occluded person re-identification (re-id), the interactive video object segmentation (VOS) and the few-shot image classification. Concretely, for the occluded person re-id task, partial features are learned by partitioning the global feature map extracted by neural networks. Pose keypoints are adopted to indicate the visible and occluded parts. The information of occluded parts is depressed. For the interactive VOS, the partial similarity between adjacent frames is important to propagate segmented masks from the previous frame to the current processing frame. Thus, the pixel distances in a local part and the global map are computed for generating masks. For the few-shot image classification, a metric-based Bayesian framework is proposed for generating robust representations and reasoning about uncertainty, including calibration, recognition of out-of-distribution images and robustness against attacks.

In sum, I investigate the significance of the discriminative and robust partial representations and the ability of estimating uncertainty for the deep learning models, and apply them to some common vision applications to illustrate the effectiveness of the deep models.

Table of contents

List of figures

Chapter 1

Introduction

1.1 Background

Deep learning models [49] have achieved remarkable success for computer vision applications. Deep learning models, which are based on artificial neural networks (*e.g.*, Convolutional Neural Networks), employ multi-layer architectures to represent multi-level features for visual imagery. Early works usually extract global representations by neural networks for vision tasks. For instance, for the person re-identification task, early deep learning models [34, 114] extract global features of person images and compute global features' distances for person retrieval. However, visual imagery (image or video) usually contains explicit parts in visual objects. There is strong psychological evidence that human intelligence recognizes visual scenes by part-whole hierarchical modeling [35]. For deep learning models, encoding the high-level partial features explicitly is an intuitive idea to obtain robust and discriminative representations, since partial features provide more abundant and diverse information. Many vision applications employ part-aware models, such as person re-identification [42, 64, 76], semantic segmentation [108], 3d object detection [100], point cloud denoising [39], *etc.* This dissertation proposes part-aware deep learning models and applies them to some fundamental vision applications, *i.e.*, person re-identification (re-id) and video object segmentation (VOS).

Another widely known limitation of deep learning models is that deep models are eager for massive labeled data. However, it is hard to acquire massive annotated data for deep model training [7] in practice. A key challenge in deep learning models is how to obtain robust representations for visual imagery, especially when the data is limited. Since it is inherent that the lack of labeled data induces uncertainties, a robust deep model should precisely estimate uncertainty. Probabilistic models have

the potential to interpret data uncertainty, and show more robustness to adversarial attacks. This dissertation proposes a probabilistic model and applies it to the few-shot image classification.

1.1.1 Part-Aware Deep Models for Vision Applications

The dissertation studies part-aware models on two vision applications: *person re-identification* and *video object segmentation*. This is because leveraging partial information has obvious advantages for the two vision applications.

Person re-identification. The person re-id task aims to retrieve a probe image from a gallery of person images. Early deep learning based models [14, 34, 116] utilize global features for distance comparison and retrieval, with the classification loss [116] or metric-learning losses (the triplet loss [34] or the quadruplet loss [14]). These previous methods aim to obtain a discriminative global representation for each person image. However, a person image inherently contains partial discriminative information for person retrieval. For instance, the color or texture of lower-body clothing may contain the key information to identify a person. Thus, learning partial features is an effective method for improving person retrieval results. Part-aware models for person re-id can be roughly classified into the following types.

Attention-based models. Attention mechanism [58, 110] is applied to the person re-id to obtain partial features directly, without extra supervisions or guidance (pose or mask). Attention-based models [58, 110] usually learn attention maps by convolutional layers, which aim at detecting different body parts. A concatenation of all learned body-part representations is used for person retrieval. Visualization results show that the learned attention regions lie in the human body parts. This is reasonable since the body parts are informative for the person re-id task.

Pose/mask-guided models. Some human-body understanding models can assist person re-id tasks, *e.g.*, human pose estimation [9, 10, 17] and human parsing [50]. Human pose estimation aims to localize human keypoints (elbows, wrists, *etc.*) in images or videos, while human parsing models provide semantic masks for body parts (face, arm, leg, *etc.*). Thus, with the guidance of human keypoints or semantic masks, re-id models [84, 85, 119] can easily detect body parts precisely and extract corresponding partial representations, which improves person retrieval performance.

Spatial partition models. For the person re-id task, prior knowledge is available that the horizontal order of a person is unambiguous; thus, horizontally partition of the global feature map can learn partial features with inherent alignment. Some approaches [64, 65, 88] conduct horizontal partition on feature maps for acquiring

partial features, instead of explicitly partitioning the person images. Compared with attention-based methods or pose/mask-guided methods, the partition models adopt a simple way to learn partial features.

The dissertation investigates the occluded person re-id, which is particularly challenging for person search in real-world applications. Owing to containing occlusions for the person images, some body parts of the person are invisible. Thus, utilizing the global representation will introduce distractive information of occlusions. In this dissertation, I propose to uniformly partition the global feature for learning partial features, and use human keypoints to indicate if one part is visible. The visible partial features are utilized to calculate the distance while the invisible parts are filtered out.

Video object segmentation. For another vision application, video object segmentation, part-aware models also play a regulatory part in boosting the quality of the segmentation masks. Video object segmentation is an essential task in the visual recognition area, whose purpose is harvesting some particular objects in all frames of one clip given the annotation or scribbles of one frame. The key challenge of VOS is how to transfer the annotated object information to other frames. Early approaches [6, 61] rely on fine-tuning with the annotated frame during evaluation, without considering the partial information of the target objects. Another limitation of the fine-tuning based methods is the slow inference speed.

Recently, some approaches [15, 38, 101] directly propagate annotated information from one frame to other frames without fine-tuning, by two-stream neural networks [101] or pixel embedding matching [15, 38]. These methods are more efficient than fine-tuning based methods. However, they still consider the target object as a whole and ignore the intra-object diversity. For instance, the early embedding-matching methods [15, 38] use each pixel of the target frame to match all the annotated object pixels, and adopt the label of the nearest neighbour. However, the local parts of adjacent frames show more similarity and are easier to propagate the label information. For example, if the target object is a person, the pixels of the head in one frame is similar to the head in adjacent frames, and dissimilar to the shoes, although the head and shoes are both parts of the target object.

This dissertation proposes a part-aware deep model for the interactive VOS. The proposed approach is pixel embedding based. It considers the diversity in one object and learns the matching distance of pixels in local parts. The part-aware model takes advantage of the pixel similarity in object local parts between adjacent frames and improves the segmentation performance.

1.1.2 Robustness and Uncertainty in Deep Learning Models

The dissertation studies the robustness and uncertainty in deep learning approaches on the few-shot image classification. The lack of labeled data for a new task inherently induces uncertainties for the few-shot learning tasks. Thus, a robust model should reason about predictive uncertainty precisely, which measures the confidence of a prediction. Higher confidence should be obtained when the prediction is more reliable. Traditional few-shot learning methods [19, 83] usually design complex deep learning models but ignore to estimate the uncertainty. Bayesian methods have the potential to undertake the above issues. Instead of point estimations for the network parameters, Bayesian methods estimate posterior distributions over model parameters.

Previous Bayesian methods [20, 77, 104] for the few-shot image classification are mostly based on MAML [19], an optimization-based approach. There exist meta parameters and task-specific parameters in these methods. Thus, they usually need multi-step optimizations for each task during evaluation, which makes the inference process time-consuming in both computation and memory.

This dissertation proposes a simple Bayesian framework for the few-shot learning, which directly models the metric between query and support set. Thus, there is no need to optimize task-specific parameters. A natural gradient based variational inference strategy is employed to approximate the distribution over model parameters. The proposed Bayesian method is effective on uncertainty estimation and obtains more robust image representations against adversarial attacks.

1.2 Thesis Organization

This dissertation is organized as follows:

- *Chapter 2*: This chapter presents a survey of methods about above mentioned vision applications, including person re-id, interactive VOS and few-shot image classification.

- *Chapter 3*: In this chapter, a part-aware deep model is proposed for the occluded re-id task. The feature map from the network is uniformly split for learning partial representations. A pose estimator is employed to extract the keypoints of person images, which indicate if a part is visible or not. Only the visible parts are used for distance computation. This work has been published at the International Conference on Computer Vision 2019 [64] and the IEEE Transactions on Neural Networks and Learning Systems [65].

- *Chapter 4*: This chapter describes a part-aware model for the interactive video object segmentation. First, pixel embeddings are extracted for each frame. Then the pixel matching maps in a local part and a whole object between frames are learned for propagating information from annotated frames to other frames. This work has been published at the Conference on Computer Vision and Pattern Recognition 2020 [63].

- *Chapter 5*: This chapter illustrates a Bayesian framework for the few-shot image classification. The proposed Bayesian model can estimate uncertainty precisely and improve robustness. The network parameters' distribution is approximated by a variational inference using the optimization of natural gradients.

- *Chapter 6*: This chapter summarizes the thesis contents and recommends future works.

Chapter 2

Literature Review

2.1 Part-Aware Deep Models for Person Re-Identification

Deep Person Re-ID. Person re-identification (re-id) is an essential problem for computer vision, whose purpose is retrieving a person by a probe from gallery images. Early works for person re-id utilize hand-crafted features [54, 62]. Thanks to the growth of deep learning, especially Convolutional Neural Networks, many deep learning-based approaches [51, 57, 64, 65, 87, 88, 97] have been proposed and achieve significantly superior performance than hand-crafted methods. For early deep models, a global representation is harvested by the identification loss [114] or metric learning losses (the triplet [34] or quadruplet loss [14]), and used for computing the distance between probe images and gallery images.

For extracting more discriminative features, some recent approaches [21, 42, 58, 88, 94, 110] have been proposed to learn partial features. For instance, some works [58, 110] employed the attention method for learning pare-aware models. Attentive deep features for different body parts were learned without extra supervisions. kalayeh *et al*. [42] employed human parsing results to extract partial body features. The final representation of the person image was constituted by partial features. Suh *et al*. [85] utilized pose estimation results for obtaining partial representations. Another group of methods for learning part-aware features is directly partitioning the global feature maps. Sun *et al*. [88] firstly partitioned the feature map horizontally and obtained the state-of-the-art performance. Following Sun [88], some approaches [21, 94] proposed to split multiple parts of part-aware features and further boosted the re-id accuracy.

Partial Person Re-ID. Occlusion is an essential difficulty for the person search in real-world applications. Early works [29, 30, 86, 113] assumed that only the probe images have occlusions and the gallery has non-occluded images. Then the probe images

were artificially cut and the non-occluded parts of the person images were persisted as the new probes. This setting is named as the partial re-id task, whose purpose is retrieving a person utilizing a partial image from holistic gallery images. The first work [113] defined the partial re-id setting, and proposed a comparison strategy of part and whole to address the task. Another group *et al*. [29, 30] employed the convolutional neural networks to harvest representation maps for the probe and gallery images, and used the spatial feature reconstruction to match the representations. Sun *et al*. [86] employed a part-aware model to learn partial features by horizontally partitioning. A region locator was designed to indicate which area is non-occluded and generates its position. The common non-occluded area across the gallery and probe images was employed to compute distances.

Occluded Person Re-ID. Partial re-id needs a manually cropping process and is inconvenient in practice. In recent years, several occluded re-id works [31, 120] appeared, which took the occluded person images as input directly without pre-processing. Zhuo *et al*. [120] proposed to train a classifier to discriminate if the person image is occluded or not. This operation helped to learn more robust representations when meeting occlusions. He *et al*. [31] tackled the occluded re-id problem by a foreground-aware pyramid reconstruction method. The pyramid features were extracted by CNN, and then the matching scores between occluded persons were computed to reconstruct the feature maps.

The above methods for the occluded re-id task have two limitations: (1) They still suppose that only the probes have obstacles and all images of the gallery are non-occluded, which is not practical in real-world scenarios. (2) They did not consider the part-aware features, which are especially important for the occluded re-id task. This dissertation relaxes the assumption and studies the situation that both the probe and gallery have images with obstacles. This new assumption is more practical in real-world applications and there exist harder cases that an image with obstacles needs to be compared with another occluded one. Besides, I propose a part-aware deep model to learn part representations, and utilize the pose information to identify the visible region and abandon the occluded region.

2.2 Part-Aware Deep Models for Video Object Segmentation

The purpose of video object segmentation (VOS) is segmenting one or several salient objects in the entire video frames. Current approaches for the video object segmentation

can be roughly divided into three categories: unsupervised, semi-supervised and interactive VOS.

Unsupervised VOS. For the unsupervised VOS task, there is no human annotation as the guidance for the user to select target objects. Models need to detect the salient objects automatically. However, this is problematic especially when there are multiple objects. Most unsupervised VOS approaches [90, 95] learned to segment the salient objects by the object motion or appearance in one video.

Semi-supervised VOS. The semi-supervised VOS task provides the label of the first frame and the purpose is harvesting the object segmentation masks of other frames. The main challenge of semi-supervised OVS is how to transfer the annotation information from a given frame to other frames. Early works [6, 61, 93] adopted fine-tuning method for semi-supervised VOS. OSVOS [6] firstly trained a two-class segmentation model using the train set, and then fine-tuned the learned network employing the first frame during evaluation. OnAVOS [93] proposed an online adaptive method employing the instances. PReMVOS [61] integrated several models utilizing fine-tuning and fusing, which acquired the state-of-the-art mIoU accuracy. Although the fine-tuning based models achieve significant performance, they are inefficient during evaluation because of the online fine-tuning operation.

Another kind of methods directly propagate the annotated information of the first frame to other frames in a video clip without fine-tuning. These methods are more efficient when testing. This kind of methods can be classified into two types. One type of these methods is matching based [15, 38, 92, 96, 103]. PML [15] extracted pixel embeddings by a convolutional neural network and matched the pixels between the first frame and other frames by a nearest neighbour classifier. VideoMatch [38] proposed to calculate similarity score maps of matching features. Above methods only consider the global matching while ignore the discriminative feature matching in a local part. Recently, FEELVOS [92] employed CNN to extract pixel embeddings in the feature space, and then matched the pixel embeddings from both the local part and the global map. CFBI [103] extended FEELVOS by considering foreground-background integration.

Another type is propagation-based [2, 69, 101], which takes as input the ensemble of the current RGB-frame and the predicted results of the previous RGB-frame. For example, RGMP [69] employed two-stream neural networks for semi-supervised VOS. One extracted the representations of the first frame and its label map, and another extracted the representations of the ensemble of the target frame and the previous frame's mask.

Interactive VOS. Above mentioned tasks (Unsupervised VOS and Semi-supervised VOS) have limitations respectively: (1) Unsupervised VOS cannot select the interest object by the user. (2) Semi-supervised VOS needs the ground truth mask of the first frame, which is time-consuming to obtain. Besides, for both two schemes, the user has no chance to refine the predicted segmentation masks and further improve the quality of predictions. Interactive VOS solved the above limitations. For the interactive VOS, a user can provide user-friendly annotations, *e.g.*, scribbles, to refine the predictions. Specifically, the user can draw scribbles on the false positive and true negative regions, and the model takes the scribbles as input to further improve the segmentation masks.

Recently, some interactive VOS approaches [3, 32, 53, 63, 70, 71] were well-developed. Two early works [3, 32] separated the interactive VOS task into two sub-tasks. Firstly the mask of the interactive frame is generated by user annotations, and then the generated mask is used to propagate information to other frames, similar to the semi-supervised VOS. Oh *et al*. [71] used two networks for the interaction and propagation, and the middle features from two networks are connected to exchange information. Above methods ignore the partial information of the objects. In this thesis, I propose to leverage the pixel matching in the local part to improve the propagation precision between frames.

2.3 Probabilistic Deep Models for Few-shot Image Classification

Few-shot Learning. In recent years, few-shot learning has attracted much attention and many few-shot learning methods [19, 22, 83, 89, 91] have been developed. The few-shot learning approaches are usually applied to the image classification, and can be divided into two types. One type is optimization-based, and the most famous method of this type is MAML [19]. MAML contains meta parameters and the meta parameters are optimized to produce task-specific parameters during inference. Since there is an optimization operation when testing, MAML is time-consuming.

Another type of the few-shot learning approach is metric-based. The metric-based methods aim at learning a proper embedding space with embedding networks, where the distance of the same class examples is close while the distance of different classes is far away. Prototypical Networks [83] employed an embedding network to learn prototypes for each class, and utilize Euclidean distance metrics to calculate the distances between embeddings of the query images and prototypes. Matching Networks [91] used two separate embedding networks for query and support examples, and utilized each labeled

sample in the embedding space as reference points for classifying the query samples. In Relation Networks [89], except for a learnable feature embedding network, the distance metric is also a learnable neural network. Few-shot GNN [22] utilized graph neural networks to model the relationship between the query set and the support set. The metric-based few-shot learning methods need no task-specific optimization and are more efficient than optimization-based methods.

Bayesian Few-shot Learning. Few-shot learning inherently induces uncertainty because of lacking data for each task. Thus, precisely estimating uncertainty improves the robustness of the few-shot learning methods. Bayesian methods have the potential to tackle this problem. Recently, some Bayesian methods for few-shot learning have been developed [20, 26, 77, 104, 107].

PLATIPUS [20], BMAML [104] and AML [77] are Bayesian extensions on MAML [19], an optimization-based few-shot learning method. Therefore there are meta parameters and task-specific parameters in these methods. These methods used variational inference to approximate either the posterior of the task-specific parameters, or the joint posterior distribution of the task-specific parameters and the meta parameters. VERSA [26] is a general probabilistic framework that used a feature extractor with the point estimation of parameters and the last fully-connection network with the distribution estimation. VFS [107] assumed that the feature embeddings were under the distribution of multivariate Gaussian. Different from previous Bayesian models, this dissertation introduces a Bayesian framework with metric-based few-shot learning models.

Chapter 3

Part-Aware Feature Learning for Occluded Person Re-Identification

3.1 Introduction

In this chapter, I address the occlusion issue in person re-identification (re-id), and propose a keypoint-guided model to align part-aware features. Occlusions always appear in real-world applications and damnify the retrieval performance since the obstacles bring delusive information and jumble the deep models. Previous methods [1, 87, 102, 114] usually extract global representations of the holistic pedestrian images, which will inevitably introduce distractive information when occlusions exist. For example, in Fig. 3.1, a failure case of the previous models with global representations shows that when a man is sheltered from a white car, the approaches that fail to remove the occlusion information will obtain error images sheltered from the same white car by mistake.

To address the occluded re-id task, I firstly define a more practical occluded re-id setting and collect a large-scale occluded re-id benchmark, which is the first occluded re-id benchmark in which both the probe and gallery have occluded pedestrian images. Then I propose a part-aware deep model to learn partial features, and employ the pose information to indicate which part is occluded. The occluded parts are depressed and the visible parts are remained for the person retrieval.

Concretely, early works [29, 30, 86, 113] for tackling the occlusion problem assumes that the probe set has obstacles and the gallery set has no obstacles. These methods artificially cut the probe images and hold the non-occluded part as the processed probes. This setting is named the *partial person re-id* problem. Therefore, the purpose of the partial re-id is retrieving one particular person in holistic impression with a

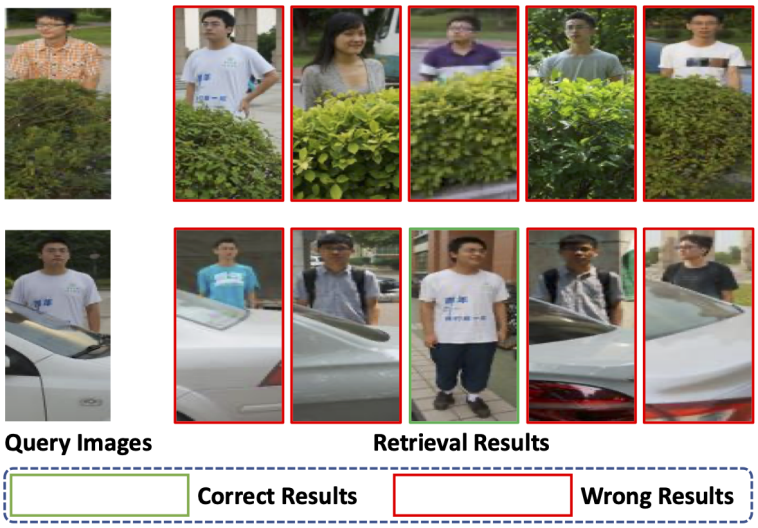

Fig. 3.1 Method [88] with global representations tends to generate error results. Images with green boundary are correct result while with red boundary are error results.

partial query as input. This partial re-id setting has two limitations: (1) In real-world scenarios, the assumption that only probe set has obstacles does not always hold. (2) This setting needs the manually cropping pre-processing which is inefficient in practice. Thus, we introduce the *occluded person re-id* setting, supposing that both the query and gallery set have obstacles. This new assumption is more practical and needs no pre-processing in real-world scenarios. We compare the partial and occluded re-id settings in Fig. 3.2. Since there is no dataset under our new assumption, we build a large-scale dataset **Occluded-Duke**, in which all images in the probe set have obstacles and the gallery images contain both occluded and holistic images.

A key solution for the occluded re-id is depressing the confusing information introduced from obstacles. We utilize the properties of the deep features from both spatial and channel perspectives to remove the obstacle information and generate the non-occluded part-aware features. In the spatial dimension, the deep feature map contains the information of the visible body parts and obstacles placed on the homologous positions [106]. Therefore, spatial attentive maps have the ability of removing the distractive information introduced by obstacles. In the channel dimension, motivated by SENet [37], channels of the deep feature map contain different responses of the object person and obstacles. Thus, adaptively balancing the channel responses can help to enhance the representations of the obstacle-free regions and depress the invisible parts. Specifically, We propose to extract pose keypoints by a pose estimator

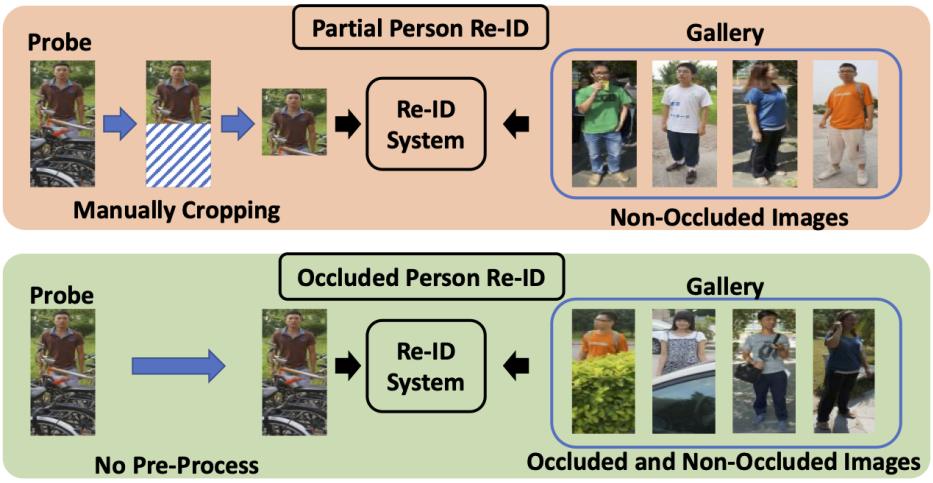

Fig. 3.2 Difference between the partial (above) and occluded re-id settings (below). The partial re-id consists of the query set with obstacles and the gallery set with non-occluded images. The query images need to be cropped to remove the occluded parts. The occluded re-id's query set has obstacles and the gallery set has both obstacles and obstacle-free images. No pre-processing like manually cropping is needed.

pre-trained on COCO [56], and use extracted keypoints to boost the re-id model in the next three perspectives.

First, the pose keypoints can indicate the visible and invisible locations of the body parts by the predictive confidence score. If the confidence of one keypoint is lower than a threshold, the corresponding body part is regarded as the occluded part. Thus, the spatial attentive maps are produced by the non-occluded keypoints and filter the deep feature map to remove the obstacle-part. This branch is named as *Keypoint-Filtered Feature* Branch of our proposed model.

Second, the human keypoints indicate if a body part is occluded or not. Thus, the landmarks can construct an informative vector to embed the occlusion-aware embeddings. Since the channels of the extracted feature map contain both the occlusion and visible information, the occlusion-aware embeddings can be used to automatically re-weight channel activations of the feature map. The re-weighted feature depresses the occlusion channel responses and enhances the visible channel responses. The operation of re-calibrating the channel activations by the keypoint embeddings is *Keypoint-Embedded Feature* Branch.

Finally, the part-aware features are generated by uniformly partitioning the feature map. During the inference time, the pose keypoints indicate which part is occluded for images in the probe and gallery sets. Thus, the part-aware features from the *common*

Table 3.1 Differences of three benchmarks on the occluded re-id.

Dataset	training set		Gallery Set		Query Set	
	Identity	Image	Identity	Image	Identity	Image
Partial-iLIDS [112]	-	-	119	119	119	119
Partial-REID [113]	-	-	60	300	60	300
Occluded-Duke	702	15,618	1,110	17,661	519	2,210

non-occluded region are selected by keypoints information and used to compute the distances when testing. In this way, the impact of the occlusions is further reduced.

We conduct generous experiments on occluded, partial, and holistic person re-id benchmarks. Our approach shows the superiority over previous re-id works [29, 30, 86, 113] significantly on both occluded and partial benchmarks, while achieving competitive or better performance on the holistic re-id benchmarks.

The following lists the main contributions of this chapter:

• An occluded re-id dataset is built, namely Occluded-Duke. It is large-scale and practical for the occluded re-id task.

• We design a pare-aware deep model for the occluded re-id, leveraging deep features' characters in both the spatial and channel dimension.

• We design a specific module to spatially depress the distractive obstacle regions in feature maps.

• We design a module to produce more discriminative features by re-weighting the channel activations utilizing the non-occluded keypoints.

• We design a matching strategy to compute the distance on the common non-occluded region between images of the probe and gallery sets when evaluating.

3.2 Occluded-Duke Dataset

Under our new assumption that both the probe and gallery set have pedestrian images with obstacles, we propose an occluded re-id setting. However, there is no appropriate dataset for the occluded re-id. Thus, a large-scale occluded benchmark is collected, called Occluded-Duke, re-split from DukeMTMC-reID [80, 115].

3.2.1 Properties of Occluded-Duke

Previous re-id datasets [112, 113, 120] for the occlusion problem suppose that all the probe set has images with obstacles but the images in gallery set are holistic. We relax this strong assumption and tackle a more practical situation: both the probe and gallery

images include occlusions. In our dataset, similar to previous datasets [112, 113, 120], all the images in the probe set are occluded to make sure that there exists at least one occluded image when evaluating. The gallery images contain both occluded and holistic images, which is practical in real-word applications. Compared with previous partial re-id datasets, our dataset is more challenging and practical. The comparison between our dataset and the partial re-id datasets [112, 113] is shown in Table. 3.1. Our dataset is much larger than early occlusion datasets. For instance, early partial datasets [112, 113] only contain hundreds of images. The Occluded-Duke has more than 35 thousands of pedestrian images, about more than 50 times larger than partial re-id datasets. Besides, the variations of our dataset are significant in both viewpoints and occlusions, including bikes, trunks, umbrellas, *etc.*, as shown in Fig. 3.3.

Fig. 3.3 Examples of the variations for occlusions.

3.2.2 Collection of Occluded-Duke

Our dataset is derived from DukeMTMC-reID [80, 115] by manual selection. The training set of our dataset is manually selected from the original DukeMTMC-reID by filtering out the images with the same occlusions in the probe or gallery set. This is

because that if there exists identical occlusions in the train and the test sets, a deep model will remember the occlusions and overfit this dataset. Thus, the generalization of the models trained on our dataset is not guaranteed. We remove totally 934 images with the same obstacles from DukeMTMC-reID to tackle the above issue.

The original DukeMTMC-reID has 14%/15%/10% occlusion pedestrian images in the train/query/gallery set. Therefore, it is inapplicable to testing the occluded re-id methods using the original DukeMTMC-reID. We collect the probe set of our new dataset by manually selecting the images with obstacles from the probe and gallery set in DukeMTMC-reID.

For the gallery set of our dataset, the original gallery set of DukeMTMC-reID is used as our gallery set, since it consists of both occluded (10%) and holistic pedestrian images (90%). Thus, there are the same images in both probe and gallery sets. Nevertheless, since the images with the same camera are not counted when testing, there is no worry that the probe image retrieves the identical image from the gallery set.

3.3 Keypoint-Guided Part-Aware Feature Alignment Model

We propose a part-aware model to address the occluded re-id problem. The key idea for addressing the occluded re-id is to detect the area with obstacles and remain the visible parts. We employ a pose estimator to extract human keypoints, which point out where the occlusions locate. We observe that if the keypoints are in the occluded region, the corresponding confidence score is lower than the visible keypoints. Thus, the visible keypoints are detected by setting a threshold to the confidence score.

The proposed part-aware model for the occluded re-id is enhanced by visible keypoints in three aspects. First, we use a module with masked features to depress the obstacles' region in the feature map spatially. Concretely, we employ non-occluded keypoints to produce Gaussian masks whose centers are located on the visible keypoints' locations. The Gaussian masks are utilized as the attentive maps to remove the occluded regions in the feature map. Second, since the visible keypoints contain the knowledge of occlusions, a keypoint-embedding can be generated by visible keypoints. We propose to automatically adjust the channel activations of the deep feature map by the keypoint-embedding. The channels contain the visible parts are enhanced while the channels contain occlusions are depressed. Thus, the re-calibrated feature map is more discriminative and robust to occlusion situations. Third, the extracted feature map is horizontally partitioned to produce part-aware features. The non-occluded pose

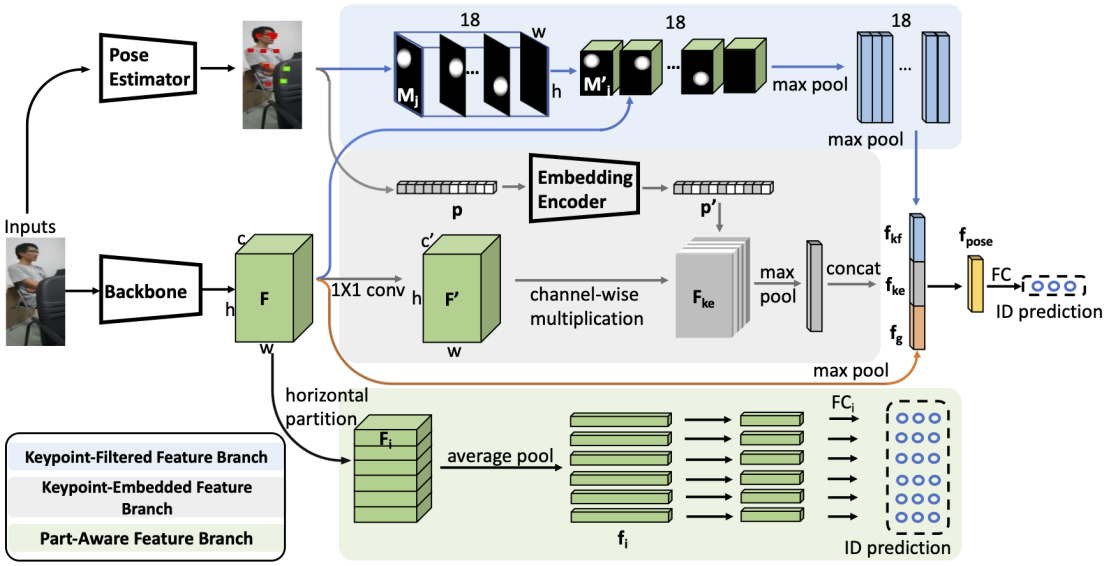

Fig. 3.4 The pipeline of our approach. We utilize red points to denote the *non-occluded* keypoints while green ones as the *occluded* keypoints. The proposed approach consists of three components. The Keypoint-Filtered Feature Branch utilizes the attentive maps generated by visible keypoints to remove occluded regions. The Keypoint-Embedded Feature Branch uses the non-occluded keypoints to produce the keypoint-embedding, which is utilized to re-weight the channel activations of the global feature map. The Part-Aware Feature Branch horizontally partitions the deep feature map to obtain the part-aware features.

keypoints provide the information that which parts are invisible during the inference time. Thus, we use part-aware features located in the common non-occluded region for comparison of the probes and gallery images.

3.3.1 Preliminaries

Fig. 3.4 illustrates the pipeline of our method, which consists of three components, $i, e,$, the Keypoint-Filtered Feature Branch, the Keypoint-Embedded Feature Branch, and the Part-Aware Feature Branch. We employ ResNet-50 [28] as our backbone, and modify it for the person re-id task following previous re-id methods [86, 88]. Specifically, the last fully connected layers of ResNet-50 [28] are removed and the size of the feature map is amplified spatially. Denote the spatial size of the input image I as $H \times W$, our modified backbone produces the feature map with 2 times larger than the original backbone. The modification is changing the stride of the fourth resblock from 2 to 1 [21, 88]. We denote the extracted deep feature map as $\mathbf{F}$.

3.3.2 Visible Keypoints Generation

As shown in Fig. 3.4, taking as input a pedestrian image I, a pose estimator pre-trained on COCO [56] is employed to generate the pose keypoints. The number of the pose keypoints is denoted as N and in this dissertation we generate eighteen keypoints including eges, legs, elbows, *etc.* For each pose keypoint, the output is the coordinates and a corresponding confidence score. By observation we found that when a keypoint is located in the occluded region, the corresponding confidence score is much lower than the visible keypoints. Thus, we utilize a threshold θ to filter out the keypoints in the occluded region and remain the visible pose keypoints. [1] The non-occluded keypoints have significant information on the obstacles from pedestrian images.

In Keypoint-Filtered Feature Branch, the non-occluded keypoints' positions are used to generate the spatial Gaussian masks. The Gaussian masks are used as the attentive masks to make the model attend to the visible body parts. Specifically, the positions of the keypoints are

$$\mathbf{P}_j = \begin{cases} (cx_j, cy_j) & \text{if } S_j^{conf} \geq \theta \\ 0 & \text{else} \end{cases} \quad (j = 1, ..., N) \tag{3.1}$$

$\mathbf{P}_j$ means the jth keypoint position. cx_j,cy_j mean the coordinate of the jth keypoint. S_j^{conf} denotes the confidence and θ denote the threshold. Therefore, we acquire non-occluded keypoints $\mathbf{P}$ with positions.

For the Keypoint-Embedded Feature Branch, the keypoint-embedding is generated for re-calibrating the channel responses. The keypoint-embedding is generated by a visible keypoint vector, which contains the significant information of obstacles. The visible keypoint vector $\mathbf{p} \in \{0,1\}^N$ is produced by

$$\mathbf{p}_j = \begin{cases} 1 & \text{if } S_j^{conf} \geq \theta \\ 0 & \text{else.} \end{cases} \tag{3.2}$$

S_j^{conf} denotes the confidence score and θ denote the threshold. Each element $\mathbf{p}_j$ of the non-occluded keypoint vector $\mathbf{p}$ means whether the jth keypoint is visible. The non-occluded keypoint vector encodes the occlusion information, which is used to generate the keypoint-embedding.

[1] If there is a case that two persons are occluded by each other, the person with more visible keypoints is treated as the target.

3.3.3 Keypoint-Filtered Feature Branch

The final features for retrieval consist of the keypoint-guided (KG) feature and the part-aware (PA) features, as shown in Fig. 3.4.

The KG feature is assembled by three parts, *i.e.*, the keypoint-filtered (KF), the keypoint-embedded (KE) and the global max-pooling feature of the global map $\mathbf{F}$.

The Keypoint-Filtered Feature Branch generates Gaussian masks by the visible keypoints as the attentive maps. In Section 3.3.2, we introduced how to obtain the locations of the visible keypoints $\mathbf{P}$. For each visible keypoint $\mathbf{P}_j = (cx_j, cy_j)$, the corresponding Gaussian mask is generated with the center at $\mathbf{P}_j = (cx_j, cy_j)$. For the occluded keypoints, we fill the mask $\mathbf{P}_j$ with $\mathbf{0}$. $\mathbf{M}_j$ is employed to represent each Gaussian mask. Since these Gaussian masks are with the same size of the input image I, we downsample the masks to the identical size as the extracted global feature $\mathbf{F}$ by bilinear interpolation. Each spatial mask $\mathbf{M}_j$ multiply the global feature map $\mathbf{F}$ to produce keypoint-filtered feature maps $\mathbf{M}'_j$, which attend to visible body parts and ignore the invisible locations.

As shown in Fig. 3.4, after generating the keypoint-filtered feature maps, the max-pooling operation is used to produce N keypoint-filtered vectors. The keypoint-filtered vectors are max-pooled to produce the KF feature $\mathbf{f}_{kf}$. We use the max-pooling operation rather than the average pooling because some visible keypoints are close to each other, and the corresponding attention masks are overlapped. Using the max-pooling operation ignores occlusions and the redundant visible body parts.

3.3.4 Keypoint-Embedded Feature Branch

The Keypoint-Embedded Feature Branch aims to re-weight the channel activations of the feature map $\mathbf{F}$. In Section 3.3.2, we introduced the visible keypoint vector $\mathbf{p}$ which encodes the obstacles' information from the pedestrian image. The visible keypoint vector $\mathbf{p}$ is used to produce the keypoint-embedding, which is the gates to re-weight the feature map in the channel dimension. However, the global feature $\mathbf{F}$ has much larger dimension compared with $\mathbf{p}$ ($2,048$ *v.s.* 18). Directly encoding the keypoint-embedding by $\mathbf{p}$ is improper. Thus, we firstly downsample the global feature map $\mathbf{F}$ to half of the original dimension (from $2,048$ to $1,024$) using a 1×1 convolutional layer. Denote the downsampled feature map as $\mathbf{F}'$. Then we utilize a two-layer network to encode the visible keypoint vector $\mathbf{p}$ to a keypoint-embedding $\mathbf{p}'$ with a dimension of $1,024$. The keypoint-embedding $\mathbf{p}'$ is treated as weights to re-weight the downsampled feature map as $\mathbf{F}'$ by channel-wise multiplication. Denote the generated keypoint-embedded

feature map as $\mathbf{F}_{ke}$. Finally, $\mathbf{F}_{ke}$ is max-pooled into the keypoint-embedded feature $\mathbf{f}_{ke}$, which contains the obstacles' information.

Obstacles Synthesis during Training. For our dataset, both the probe and gallery pedestrian images have a large variety of occlusions. Nevertheless, the training set contains only a small quantity of occluded pedestrian images. Thus, a gap between the training process and the testing appeared. The varieties of occlusions in the training set are not enough to generate different kinds of non-occluded keypoint vectors. Thus, we synthesize occluded images for the training set by randomly erasing the training images to enlarge the variety of occlusions. The keypoints in the erased region are treated as the occluded keypoints.

3.3.5 Objective Function

The final keypoint-guided feature is concatenated by three parts, *i.e.*, the keypoint-filtered feature $\mathbf{f}_{kf}$ in the Keypoint-Filtered Branch, the keypoint-embedded feature $\mathbf{f}_{ke}$ in the Keypoint-Embedded Branch, and the global max-pooling feature $\mathbf{f}_g$ of the feature map $\mathbf{F}$. This dimension of the assembled feature is reduced to 256 utilizing a FC layer. Denote this new vector as the keypoint-guided feature $\mathbf{f}_{pose}$. A classification loss is adopted which takes as input $\mathbf{f}_{pose}$. The loss function of the branches for the keypoint-guided feature is

$$\mathcal{L}_{keypoint} = CrossEntropyLoss(\hat{z}, z)\,. \tag{3.3}$$

$\hat{z}$ and z denote the predicted class and the ground truth respectively.

Another branch of our proposed method is the Part-Aware Feature Branch, as shown in Fig. 3.4. The Part-Aware Feature Branch aims to generate part-aware features, by partitioning the global feature map $\mathbf{F}$ horizontally into the part-aware feature maps, $\mathbf{F}_i$. The total partition number is p and $i = 1, ..., p$. After that each part-aware map $\mathbf{F}_i$ is averaged to produce the part-aware vector $\mathbf{f}_i$. The dimension fo the part-aware vector $\mathbf{f}_i$ is reduced to 256. A classification loss is employed with the input of $\mathbf{f}_i$. Therefore, the loss for the part-aware Branch $\mathcal{L}_{part}$ is

$$\mathcal{L}_{part} = \sum_{i=1}^{p} CrossEntropyLoss(\hat{z}_i, z). \tag{3.4}$$

$\hat{z}_i$ is the identity predicted by the i-th feature. The final loss $\mathcal{L}$ consists of the above losses,

$$\mathcal{L} = \alpha\mathcal{L}_{part} + (1-\alpha)\mathcal{L}_{keypoint}\,. \tag{3.5}$$

α is a balancing coefficient between $\mathcal{L}_{part}$ and $\mathcal{L}_{keypoint}$.

3.3.6 Part-Aware Feature Matching in Shared Visible Region

During the evaluation, for each pedestrian image we have a keypoint-guided feature and p part-aware features. These part-aware features contain both visible and invisible features. Thus, we employ the keypoints to instruct which part-aware feature is occluded. A Part-Aware Feature Matching Strategy is shown in Fig. 3.5. If there exist visible keypoints in one part, the corresponding part-aware feature is viewed as the visible feature, and vice versa. Thus, the distances across the probe and gallery pedestrian images are computed by the visible part-aware features in the shared non-occluded region. Besides, the keypoint-guided feature is also utilized for retrieval. The above distances computed by non-occluded part-aware features and the keypoint-guided feature are averaged to generate the final distance.

Concretely, a non-occluded mark $l_i \in \{0, 1\}$ for each part-aware feature $\mathbf{f}_i$ is decided by the visible keypoints. $l_i = 0$ denotes the part-aware feature is visible while $l_i = 0$ denotes the part-aware feature is occluded. Thus,

$$l_i = \begin{cases} 1 & \text{if } \exists cy_j \in [\frac{i-1}{p}H, \frac{i}{p}H) \\ 0 & \text{else} \end{cases} \quad (j = 1, ..., N)\,. \tag{3.6}$$

H is the height of the input pedestrian image while cy_j denotes the jth longitudinal coordinate of the keypoint $\mathbf{P}_j$.

For the ith part, the distance across the probe and gallery is computed by

$$d_i = Dist(\mathbf{f}_i^p, \mathbf{f}_i^g) \quad (i = 1, ..., p)\,. \tag{3.7}$$

$\mathbf{f}_i^p$ denotes the ith part-aware vector of the query while $\mathbf{f}_i^g$ denotes the part-aware vector of gallery. $Dist(\cdot)$ denotes a cosine distance function.

For the distance computed by keypoint-guided features,

$$d_{pose} = Dist(\mathbf{f}_{pose}^p, \mathbf{f}_{pose}^g)\,, \tag{3.8}$$

where $\mathbf{f}_{pose}^p$ denotes the keypoint-guided feature of the probe while $\mathbf{f}_{pose}^g$denotes the keypoint-guided feature of the gallery, respectively.

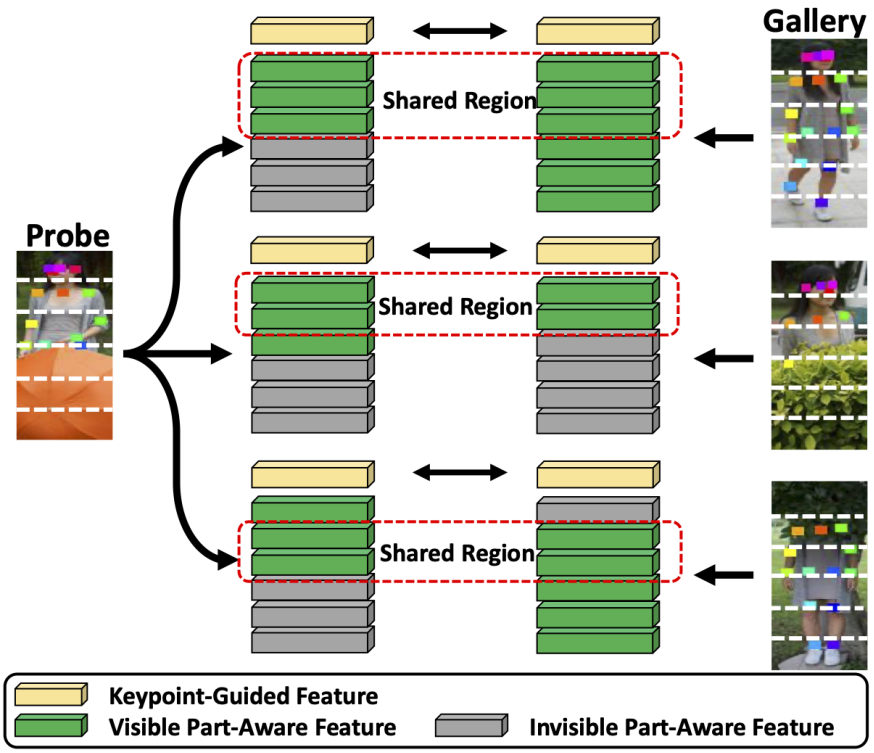

Fig. 3.5 Distance comparison strategy of our approach. The distances across images from probe and gallery sets are computed using part-aware features in the common non-occluded region as well as the keypoint-guided feature.

The final distance is the average of the distances calculated by the part-aware features in the shared non-occluded region and the keypoint-guided vector.

$$dis = \frac{\sum_{i=1}^{p}(l_i^p \cdot l_i^g)d_i + d_{pose}}{\sum_{i=1}^{p} l_i^p \cdot l_i^g + 1}. \tag{3.9}$$

dis means the final distance. l_i^p and l_i^g are the ith non-occluded mark of the query and gallery, respectively.

3.4 Experiments and Analysis

3.4.1 Datasets

We evaluate our method on one occluded, two partial, and two holistic re-id datasets.

Occluded-Duke is our proposed dataset specifically built for the occluded re-id. This dataset is large-scale, including 15,618 images in the training set, 17,661 images

in the gallery set, and 2,210 images in the query set. All pedestrians in the query set are with obstacles while 10% gallery images are with obstacles.

Partial-REID [113] is a partial re-id dataset, containing 60 identities and 600 pedestrian images. In this dataset, 300 pedestrian images are with obstacles, while 300 pedestrian images are holistic. The occlusions in Partial-REID include trees, cars, desks, *etc.*

Partial-iLIDS [112] contains 238 pedestrian images and 119 pedestrian identities. For each person identity, there are an occluded pedestrian and a holistic pedestrian. For the two partial datasets, the partial methods manually cut the images with obstacles first, then remain the non-occluded parts as the new probe images. Under our setting, our approach needs no cutting pre-processing.

Market-1501 [111] are composed of 32,668 images of 1,501 person identities, including 12,936 pedestrian images for training and 19,732 pedestrian images for testing. All images are collected from six cameras. Since there are few images contain occlusions in Market-1501, it is viewed as the holistic person re-id dataset.

DukeMTMC-reID [80, 115] has 1,812 identities from eight cameras. There are 2,228 query images, 16,522 training images and 17,661 gallery images in this dataset. Since the query set only has more than 85% holistic pedestrian images and the gallery set has more than 90% holistic images, this dataset is treated as a holistic dataset.

Evaluation Metrics. For evaluating the re-id performance, we employ the commonly-used metrics as in previous re-id methods, *i.e.*, the cumulative matching cure (CMC) and the mean Average Precision (mAP).

3.4.2 Implementation Details and Hyperparameters

ResNet-50 [28] is employed as our backbone and is pre-trained by ImageNet [16]. We employ AlphaPose [17, 98] as the pose estimator, which is pre-trained on COCO [56]. The threshold θ for distinguishing the visible and invisible keypoints is 0.2. Similar to previous person re-id methods [31, 88], the resolution of inputs is adjusted to 384×128 when we train our model. During training, the batch size is set to 32 and the number of the training epochs is 60. When we train our model on our dataset and two non-occluded datasets, we set the learning rate as 0.1. We reduce the learning rate to 0.01 during the final 20 epochs. The coefficient α for balancing the losses is 0.5. When we train our model on two partial re-id datasets, the start learning rate is 0.02. Then we reduce it to 0.002 for the final 20 epochs. The coefficient α is 0.9. We utilize the data augmentations, including random erasing and random flipping. Random erasing

Table 3.2 Person Re-ID Results on the Occluded-Duke dataset.

Method	Rank-1	Rank-5	Rank-10	mAP
DIM [105]	0.215	0.361	0.428	0.144
LOMO+XQDA [54]	0.081	0.17	0.22	0.05
Part Aligned [110]	0.288	0.446	0.51	0.202
Random Erasing [118]	0.405	0.596	0.668	0.3
HACNN [52]	0.344	0.519	0.594	0.26
Triplet [34]	0.355	0.528	0.611	0.27
Aligned reID [109]	0.415	0.588	0.657	0.327
Adver Occluded [40]	0.445	-	-	0.322
PCB [88]	0.426	0.571	0.629	0.337
Part Bilinear [85]	0.369	-	-	-
FD-GAN [23]	0.408	-	-	-
PGFA [64]	0.514	0.686	0.749	0.373
DSR [29]	0.408	0.582	0.652	0.304
SFR [30]	0.423	0.603	0.673	0.32
Ours	**0.563**	**0.724**	**0.78**	**0.435**

Table 3.3 Evaluation time on Occluded-Duke. The evaluation time is the milliseconds per query.

Method	Evaluation Time	Method	Evaluation Time
$\text{PGFA}_{\text{w/o key}}$ [64]	130ms	PCB [88]	90ms
$\text{PGFA}_{\text{w/ key}}$ [64]	780s	DSR [29]	4540ms
$\text{Ours}_{\text{w/o key}}$	140ms	SFR [30]	4760ms
$\text{Ours}_{\text{w/ key}}$	790ms	-	-

is employed to synthesize the pedestrian images with obstacles in the . The keypoints in the erased region are viewed as the invisible keypoints.

3.4.3 Evaluation Performance

Performance on Occluded-Duke. Table. 3.2 illustrates the performances of our approach and early state-of-the-arts on our proposed Occluded-Duke. The first,second,third group indicates the methods for the holistic re-id, the methods using the pose keypoints, the approaches for the partial re-id, respectively. The fouth group shows our model. Our model outperforms the previous approaches significantly.

Our part-aware feature branch, which horizontally partitioning the global feature map, is similar to PCB [88]. Differently, our method employs a matching strategy that part-aware features across the probe and gallery pedestrian images in the shared

Table 3.4 Person Re-ID Results on Partial-REID and Partial-iLIDS.

Method	Partial-REID		Partial_iLIDS	
	Rank-1	Rank-3	Rank-1	Rank-3
AMC+SWM [113]	0.373	0.46	0.21	0.328
MTRC [55]	0.237	0.273	0.177	0.261
DSR [29]	0.507	0.7	0.588	0.672
SFR [30]	0.569	0.785	0.639	0.748
VPM [86]	0.677	0.819	0.672	0.765
PGFA [64]	0.68	0.8	0.69	0.809
Ours	**0.725**	**0.83**	**0.706**	**0.813**

Table 3.5 Person Re-ID Results on Market-1501 and DukeMTMC-reID.

Method	Market-1501		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
SVDNet [87]	0.823	0.621	0.767	0.568
BoW+kissme [111]	0.444	0.208	0.251	0.122
PAN [115]	0.828	0.63	0.717	0.515
PAR [110]	0.81	0.634	-	-
Pedestrian[117]	0.82	0.63	-	-
DSR [29]	0.835	0.642	-	-
MultiLoss [51]	0.839	0.644	-	-
TripletLoss [34]	0.849	0.691	-	-
Adver Occluded [40]	0.865	0.783	0.791	0.621
MLFN [12]	0.9	0.743	0.81	0.628
PCB [88]	0.924	0.773	0.819	0.653
PGFA [64]	0.912	0.768	0.826	0.655
Ours	**0.927**	**0.813**	**0.862**	**0.726**

visible region are used for computing distances. Besides, the visible keypoints are used to generate attention masks and re-calibrate channel responses for generating the keypoint-guided feature. Considering one baseline method, PCB [88], our approach outperforms it by more than +13% Rank-1 and around +10% mAP, illustrating the significance of our approach.

Considering PGFA [64], our approach adds the keypoint-embedded feature and uses simulation augmentation using the random erasing. Table. 3.2 demonstrates that employing the keypoint-embedding boosts Rank-1 by about +5%, while boosts mAP by about +6%, illustrating the benefits of Keypoint-Embedded Feature Branch.

Considering two partial re-id approaches [29, 30], our method surpasses them by 15.5% and 14.0% Rank-1 score, respectively. Besides, our method does not need the pre-processing of manually cropping.

Table. 3.3 shows the comparison of the inference speed between our approach and PCB [88], PGFA [64] and the partial re-id approaches, *i.e.*, DSR [29] and SFR [30]. "w/ key" or "w/o key" denote our approach with or without predicting the human keypoints. Table. 3.3 demonstrates that our approach needs a little more time compared with PGFA owing to the additional Keypoint-Embedded Branch. Comparison of our method with or without keypoints prediction shows that extracting the pose keypoints is inefficient. Thus, extracting the pose keypoints for each pedestrian image in advance is practical for real-world scenarios. The partial re-id methods (DSR and SFR) are much more time-consuming than our approach because they need an inefficient feature map matching process during evaluation.

Performance on Partial Datasets. Recently, several re-id methods [29, 30, 55, 86, 113] for the partial re-id appeared. We evaluate our method and these partial re-id methods on two partial datasets, Partial-REID and Partial-iLIDS, and the comparison is shown in Table. 5.1. Our model is trained utilizing the of Market-1501, which is the same as previous partial re-id methods [29, 30, 86]. Table. 5.1 illustrates that our model outperforms previous partial re-id methods [29, 30, 55, 86, 113] by a large margin on partial datasets, without the manually cropping pre-processing.

Performance on Holistic Datasets. The performances of our approach and previous re-id methods on the holistic re-id datasets are presented in Table.3.5. The proposed model surpasses all previous state-of-the-art re-id approaches, demonstrating that the proposed approach has the ability of tackling both the occluded and holistic re-id tasks.

3.4.4 Ablation Studies

The Impact of the Keypoint-Embedded Feature Branch. The Keypoint-Embedded Feature module utilizes the non-occluded keypoint vector to produce the keypoint-embedding. The keypoint-embedding is served as weights to re-weight channel responses of the feature map. To enlarge the variety of the non-occluded keypoints, we synthesize occluded pedestrian figures by random erasing during training. In Table. 3.6, $\text{Ours}_{\text{w/o syn w/o ke}}$ indicates our model without the keypoint-embedded branch and data simulation. $\text{Ours}_{\text{w/o ke}}$ indicates our approach with no keypoint-embedded features and $\text{Ours}_{\text{w/o syn}}$ means our approach without data synthesis. Experiment results

Table 3.6 The ablation studies on the Keypoint-Embedded Feature Branch, the Keypoint-Filtered Feature Branch and the part-aware feature matching strategy.

Method	Rank-1	Rank-5	Rank-10	mAP
$\text{Ours}_{\text{w/o syn w/o ke}}$	0.514	0.686	0.749	0.373
$\text{Ours}_{\text{w/o ke}}$	0.532	0.694	0.753	0.405
$\text{Ours}_{\text{w/o syn}}$	0.541	0.696	0.752	0.403
$\text{Ours}_{\text{w/o kf}}$	0.55	0.705	0.767	0.415
$\text{Ours}_{\text{w/o matching}}$	0.512	0.624	0.734	0.412
Ours	**0.563**	**0.724**	**0.78**	**0.435**

demonstrate that adding this keypoint-embedded branch boosts the mAP performance by +3%. Adding the operation of data synthesis further boosts mAP by another +3%.

The Impact of the Keypoint-Filtered Feature Branch. The Keypoint-Filtered Branch generates attentive maps and removes the distractive information in the occluded region. In Table. 3.6, $\text{Ours}_{\text{w/o kf}}$ denotes our approach without the Keypoint-Filtered Branch. Results show that adding this branch boosts Rank-1 by +1.3% and mAP by +2%, illustrating the benefits of the Keypoint-Filtered Branch.

The Impact of Matching Strategy in Shared Visible Region. Our matching strategy utilizes the visible keypoints to point out which part-aware feature is occluded. Only the part-aware features in the shared visible region are used for computing the distances. To demonstrate the effectiveness of our matching strategy, we evaluate our model with all the part-aware features. In Table. 3.6, $\text{Ours}_{\text{w/o matching}}$ denotes our approach without the matching strategy in the shared visible region. Results show that using the proposed matching strategy boosts the Rank-1 performance by +5.0% and the mAP performance by +2.3%.

Ablations on the Coefficient α. In Equation 3.5, the final loss is composed of two components, $\mathcal{L}_{part}$ and $\mathcal{L}_{keypoint}$. $\mathcal{L}_{part}$ denotes the loss of the part-aware features and $\mathcal{L}_{keypoint}$ denotes the loss of the keypoint-guided feature. α is the coefficient to balance $\mathcal{L}_{part}$ and $\mathcal{L}_{keypoint}$. The ablation studies are conducted on α, which increases from 0 to 1. $\alpha = 0$ indicates the branches about the keypoint-guided feature are used only, and $\alpha = 1$ indicates the Part-Aware Feature Branch is used only. Fig. 3.6 (a) illustrates that the re-id results on $0 < \alpha < 1$ are better than $\alpha = 0$ or $\alpha = 1$, indicating that assembling these two losses acquires better Rank-1 or mAP than using one loss. When $\alpha = 0.5$, our approach can achieve the best result.

Ablations on the Part Number. We use the part number p to denote granularity of the part-aware features. $p = 1$ indicates we only use the global feature and no part-aware features are generated. Fig. 3.6 (b) shows that when $p > 1$, our model achieves

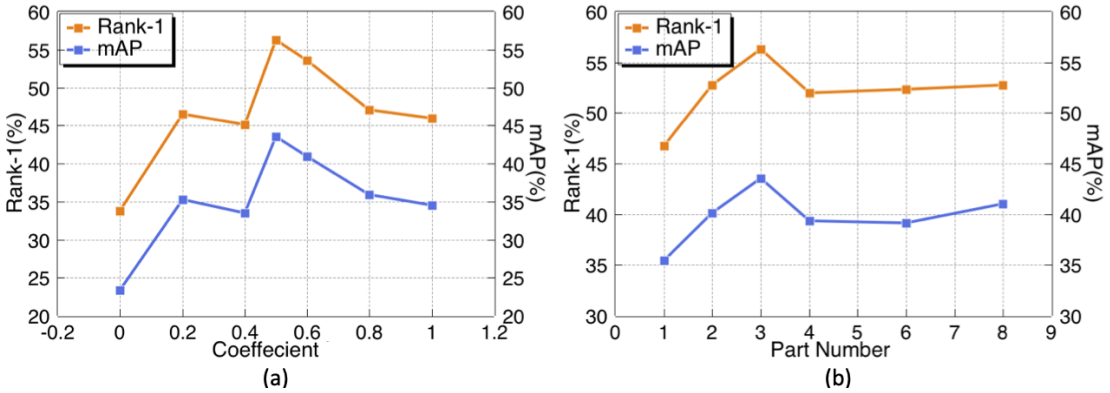

Fig. 3.6 (a) The impact of the coefficient α. When α is 0, our model utilizes the keypoint-guided feature only. When α is 1, we adopt the part-aware features for evaluation. (b) Ablations on the part number p.

Table 3.7 Ablations on multiple granularities.

Method	Rank-1	Rank-5	Rank-10	mAP
$p = 2, 3$	54.3	0.708	0.773	0.423
$p = 3, 4$	0.563	0.722	0.781	0.436
Ours ($p = 3$)	0.563	0.724	0.78	0.435

better the re-id performance than $p = 1$, demonstrating that generating part-aware features is effective on the occluded re-id task. At the point of $p = 3$, the proposed approach acquires the best performance. If $p > 3$, the performance drops because some partition region is too small and easily to be ignored because it contains no visible keypoints, although it contains visible body parts. Table. 3.7 shows the comparison across our approach and multi-granularity approaches [21, 94]. $p = 2, 3$ denotes we split the global feature to two and three parts simultaneously. Experiments demonstrate that the performance is similar to ours when utilizing multi-granularity.

Ablations on the Gaussian Heatmap. The Gaussian maps are generated by the location of the keypoints. The parameter σ of the gaussian map determines the size of the information window. Small σ means the visible region in the generated map is small. Large σ means the visible region is large and the attentive maps cannot filter out the distractive occlusions. Fig. 3.7 (a) illustrates the ablations on the hyperparameter σ. When σ equals to 20, our model has the best performance.

Fig. 3.7 (b) demonstrates the ablations on guassian map resolutions. Results show that amplifying the input resolution from 384×128 to 570×190 does not affect the re-id accuracy. When the rosolution is too small (240×80), the performance drops because the input information is limited.

Table 3.8 Ablations on different human pose estimators.

Method		Rank-1	Rank-5	mAP
AlphaPose [17]	$\text{Ours}_{\text{w/o syn w/o key}}$	0.514	0.686	0.373
	$\text{Ours}_{\text{w/o syn}}$	0.541	0.696	0.403
	Ours	0.563	0.724	0.435
OpenPose [8]	$\text{Ours}_{\text{w/o syn w/o key}}$	0.491	0.667	0.353
	$\text{Ours}_{\text{w/o syn}}$	0.523	0.676	0.385
	Ours	0.543	0.712	0.412

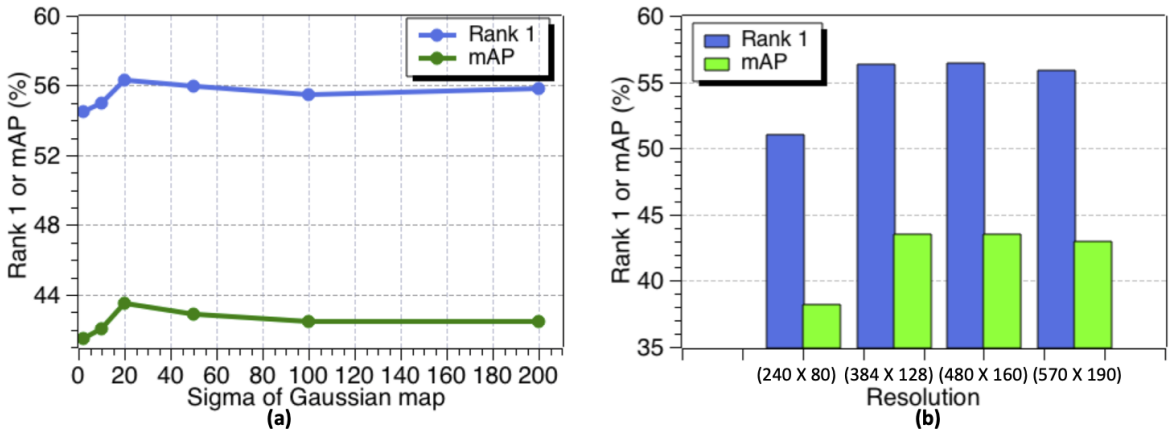

Fig. 3.7 (a) Ablations on the hyperparamters σ of Gaussian maps. (b) Ablations on the image resolutions.

The Ablations on the Pose Estimators. There are two commonly-used pose estimators, AlphaPose [17] and OpenPose [8]. The above experiments are based on AlphaPose [17]. In Table. 3.8, we compare the re-id performance of our approach between the two pose estimators, and results show that our models with the two pose estimators achieve similar performance, indicating our approach is robust to keypoint extracting models.

3.4.5 Visualization

Fig. 3.8 shows the visualization of the attentive maps. The attentive maps are produced by the visible keypoints and in Fig. 3.8 the maps attend the visible body parts in a pedestrian image, and obstacles' regions are suppressed.

Fig. 3.9 illustrates several retrieval outputs of the baseline [88] and our approach on our dataset Occluded-Duke. The PCB model uses the invisible features and cannot filter out the distractive information of occlusions, and the query image is easy to

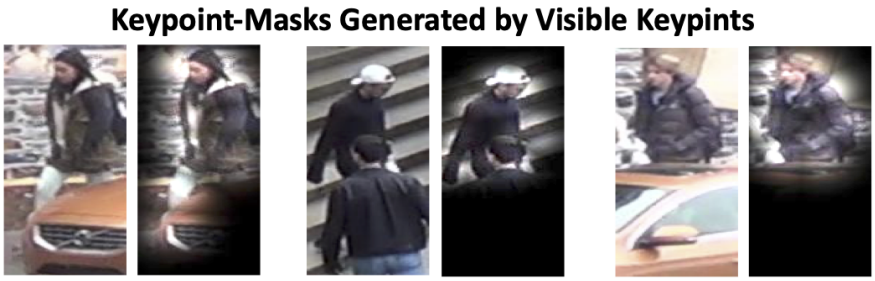

Fig. 3.8 The appearance of the keypoint-masks produced by non-occluded keypoints.

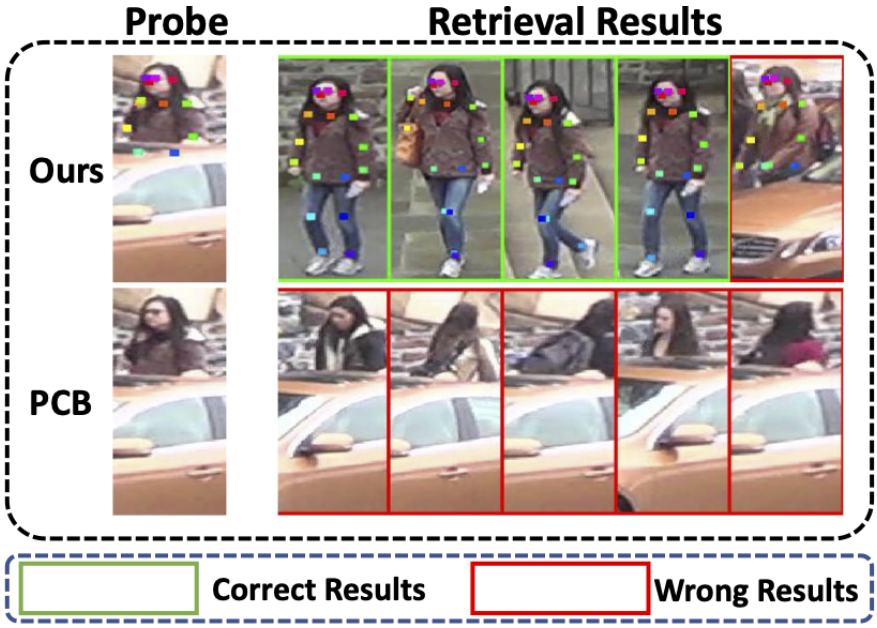

Fig. 3.9 Visulization of the outputs on the baseline [88] and the proposed approach.

retrieve error images with similar occlusions. Differently, our approach depresses the impact of the occlusions and retrieves more correct results.

3.5 Conclusion

In this chapter, we propose a part-aware deep model for the occluded re-id task. We firstly define a new setting that both the probe and gallery sets contain obstacles. Under this assumption, we build a large-scale occluded re-id benchmark, Occluded-Duke. We propose to leverage the pose keypoints to detect the occluded region and enhance the discriminative feature. First, the attentive maps are generated by locations of the pose keypoints to filter out the occlusion information; second, the visible keypoints are used to generate keypoint-embeddings, which serve as gates to re-weight the

channel responses. Third, during evaluation, the part-aware features in the common non-occluded region are employed for comparison.

Chapter 4

Part And Whole Matching for Interactive Video Object Segmentation

4.1 Introduction

This chapter presents a part-aware deep model which focuses on the interactive video object segmentation (VOS). Interactive VOS aims to acquire segmentation masks of one or several objects in a video clip when some user's annotations are provided, for example, scribbles. Specifically, Fig. 4.1 shows the turn-based interactive VOS proposed by Caelles [7]. In the first turn, the users provide annotations on the foreground objects at some frame in a video clip, and a model predicts the segmentation mask of these foreground objects at the interactive frame. Then the information of segmentation mask is propagated to other frames temporally to obtain the object masks in the entire video. In next turns, the users can draw scribbles on the false positive and true negative regions to refine the segmentation results on one frame. The user interaction and mask transference are repeated until the users are satisfied with the segmentation results. The interactive VOS is efficient and flexible than semi-supervised VOS and unsupervised VOS because it needs user-friendly annotations on a frame per turn, and is able to control the final results by repeatedly providing user annotations.

Previous approaches for the interactive VOS [3, 7, 33, 67, 70] have some limitations: (1) only the global structural information is considered for both interactive mask generation and mask transference. The scribble-labeled frame and scribbles are directly taken as the input of neural networks, to extract the global features for generating

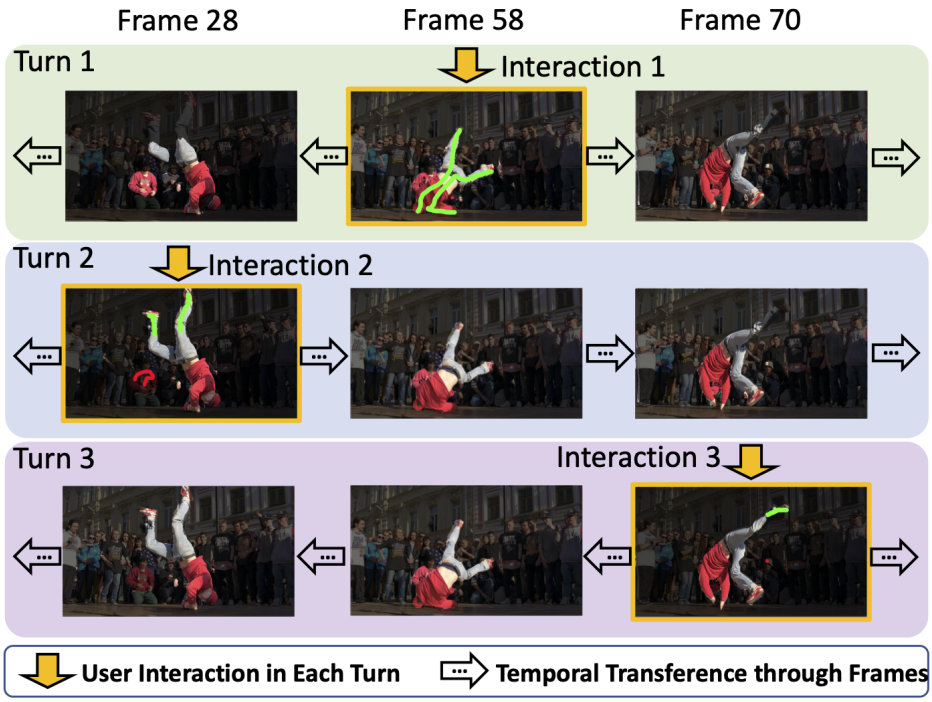

Fig. 4.1 An example of the turn-based interactive VOS. The green and red scribbles denote the true nagative and false positive regions, respectively. At the first turn, the user provides green scribbles at frame 58. The model generates the segments of objects on the annotated frame and transfers the segments to other frames. In the following turns, the user refines the results by repeatedly drawing scribbles. For instance, at the second turn, the user draws green and red scribbles at frame 28.

segmentation masks. However, partial information in local regions is important during transference because local parts between adjacent frames are similar; (2) the inference of the entire networks [3, 70] is executed for each interactive turn, which is inefficient; (3) the interaction and transference are conducted using two independent networks [3, 33]; (4) only the results of previous turns are utilized for the new turn, which ignores the information from multi-turn interactions [33].

Considering the above limitations, we propose a part-aware and unified method called Memory Aggregation Networks, which is more efficient since the features are extracted only at the first turn. The interaction and transference operations are unified into one framework, illustrating the elegance of our method comparing with previous methods.

Specifically, the interaction and transference networks are unified by sharing an embedding encoder to extract pixel embedding vectors for each frame in a video. The embedding encoder extracts an embedding vector for each pixel in a space, in which pixels located in an identical object are near to each other. After extracting pixel

vectors, our method predicts masks of user-labeled frame as well as other frames by two "shallow" segmentation heads with four convolutional layers, respectively. This is more efficient especially for the turn-based interactive VOS because we only extract pixel embedding vectors for all frames at the beginning turn. For next turns, we utilize the pixel vectors and "shallow" heads to predict segmentation masks, which is more efficient than early works. Moreover, a local part matching approach is adopted to transfer the mask information across frames precisely. Besides, we propose memory aggregation modules to accumulate and arrange informative knowledge from all previous interactive turns. The accumulated information improves the robustness of the predictive masks and boosts the accuracy significantly.

We evaluate our method on an interactive VOS benchmark at the DAVIS Challenge 2018 [7]. Results show the superiority of our approach than previous state-of-the-arts without additional Youtube-VOS dataset [99], CRF post-processing [47, 67] or the optical flow information [33]. Besides, considering the efficiency, our method completes 7-turn interactive operations in 60s while a previous state-of-the-art [70] only completes 5-turn.

4.2 Memory Aggregation Networks with Part and Whole Matching

For the turn-based interactive VOS task, given user annotations on one frame, the learned model harvests segmentation masks of the objects in the entire video clip. Users can refine the results repeatedly on one frame per turn to obtain better masks until satisfied. Most of the previous approaches [3, 33, 70] only focus on the global structural information of the frames and annotations, while ignoring the partial feature similarity of adjacent frames. This dissertation introduces a part-aware model, excavating the local part matching to precisely transfer the segmentation information through the previous to the present frame. Our model tackles the interaction and transference in a consolidated framework.

Fig. 4.2 shows the holistic framework of our method. Our method is composed of three components: a pixel vector encoder, an interaction branch, and a transference branch. This pixel vector backbone extracts the pixel embeddings for each frame in the given video and assigns each pixel a pixel embedding vector. The pixel embedding vectors construct an embedding space. The purpose of our method is to learn the pixel vectors close to each other if they are located in the identical object while distant if they are located in different objects in the embedding space. The interaction branch

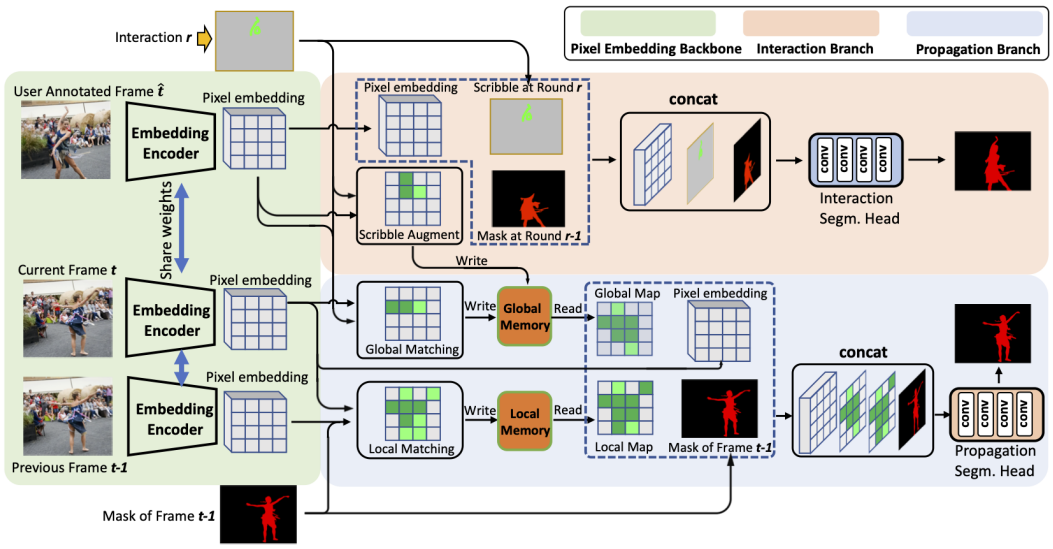

Fig. 4.2 The pipeline of our method, consisting of **pixel vector encoder**, an **interaction branch** and an **transference branch**. The pixel embedding vectors are extracted by this pixel vector encoder for each frame in one video. The interaction branch utilizes a "shallow" segmentation head to generate the segmentation results of the user-labeled frame. While the transference branch employs memory modules to accumulate the discriminative information and a segmentation head to predict segments of other frames. For the whole and partial maps, deeper green pixels have predictive results with higher confidence.

utilizes the users' annotations and the pixel vectors to generate the segmentation masks for the interactive frame. The transference branch leverages the mask information of the interactive frame as well as the previous frame to generate segment results of the present frame. The interaction branch and the transference branch utilize one identical pixel vector encoder. The pixel vectors are extracted for the entire video at the first turn. Thus, it is efficient for our method to predict the object masks in the video. In this dissertation, we denote the interactive frame where the users draw scribbles as the $\hat{i}^{th}$ frame (also called the user-annotated frame). We also denote the present frame as the i^{th} frame and the previous frame as the $(i-1)^{th}$ frame. The pixel of the present frame is denoted as p, while the pixels assigned to the object o in the interactive or the previous frame as q.

Pixel Vector Encoder. The pixel vector encoder extracts pixel embedding vectors for each frame in a video. In the learned vector space, the pixel vectors in an identical object are close to each other and in different objects are distant. We use DeepLabv3plus [13] with ResNet-101 [28] as our backbone of the pixel vector encoder. An embedding layer with a 3×3 kernel is employed after the backbone to

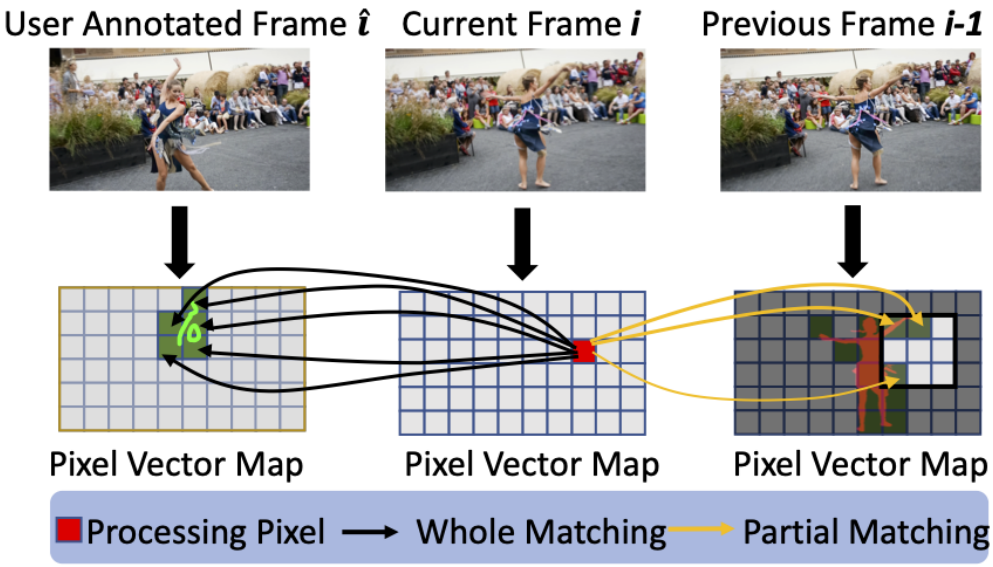

Fig. 4.3 Matching operation on the whole map and local part. With regard to a pixel p in the present frame (the i^{th} frame), we compute distances between p and pixels assigned to the object label in the interactive frame (whole map) by scribbles or the previous frame (partial map) by predicted segmentation mask. The nearest neighbor of the pixel p in the embedding space is utilized to generate the matching map.

reduce the dimension to 100. For the pixel p from the input image, we denote the corresponding pixel embedding vector as $\mathbf{v}_p$. The embedding space of the pixel vector $\mathbf{v}_p$ is an Euclidean space. Following [18, 92], we define a normalized distance between pixels p and q utilizing the corresponding pixel embeddings $\mathbf{v}_p$ and $\mathbf{v}_q$ as

$$dist(p,q) = 1 - \frac{2}{1+\exp(\|\mathbf{v}_p - \mathbf{v}_q\|_2^2)}. \tag{4.1}$$

The function normalizes the distance between p and q from 0 to 1. A smaller number means a closer distance. Following FEELVOS [92], we construct matching maps using the pixel distances to build matching maps, which are utilized by "shallow" convolutional heads to predict the final masks.

Transference Branch. The transference branch transfers the segmentation masks from the interactive $\hat{i}^{th}$ and the previous $(i-1)^{th}$ frames to the present processing i^{th} frame for predicting object masks. Similar to FEELVOS [92], the whole and partial matching is used to produce the whole and the partial map, as shown in Fig. 4.3. Differently, our method employs memory modules to store, accumulate and update the informative knowledge of matching maps. With the user-annotation turn increasing,

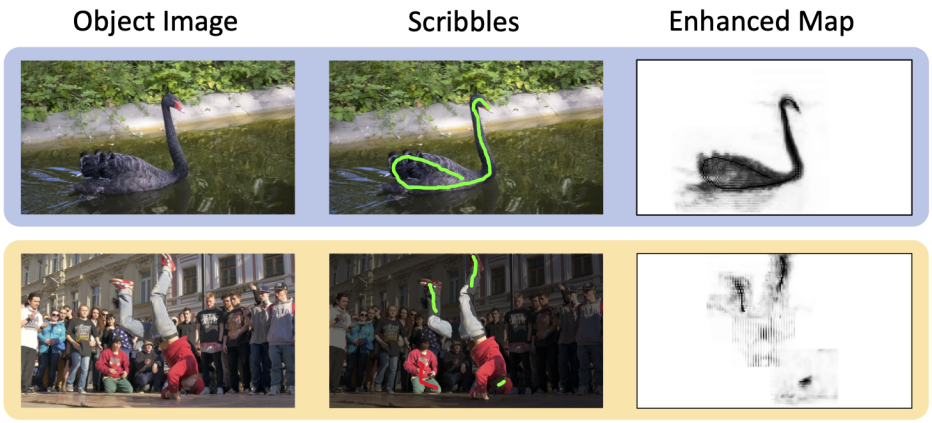

Fig. 4.4 The scribbles are enhanced by computing the distance in the pixel vector space.

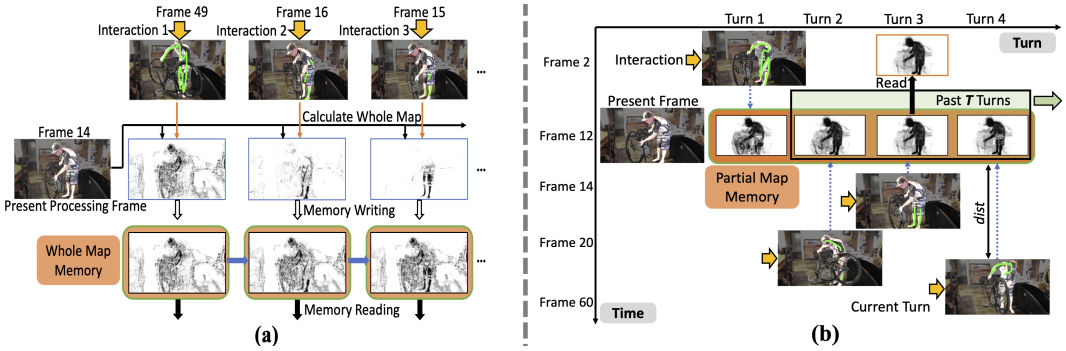

Fig. 4.5 **(a) Memory module for whole matching maps**. Whole matching maps from the transference branch is accumulated as well as updated in this memory module. **(b) Memory module of the partial matching maps and the forgetting mechanism.** The partial matching maps in the transference branch is recorded. partial maps from the past R turns are used to predict the segmentation masks.

the discriminative information stored in memories becomes more sufficient and robust to generate accurate masks. This is specially designed for the interactive VOS.

Memory Module for Whole Matching Map. We denote $\mathbb{P}_i$ as the set of pixels in the i^{th} frame. Denote $\mathbb{P}_{\hat{i},o,t}$ as the set of pixels labeled as object o in the $\hat{i}^{th}$ frame at the interactive turn t. For pixel p in the set $\mathbb{P}_i$, the whole matching map is generated by calculating the distance between p and the nearest neighbour from $\mathbb{P}_{\hat{i},o,t}$, as show in the left of Fig. 4.3. Formally, we have

$$\mathbf{W}_{i,t}(p) = \min_{q \in \mathbb{P}_{\hat{i},o,t}} dist(p,q). \tag{4.2}$$

For the semi-supervised VOS, the model is given a groud truth mask at the first frame. Differently, for the interactive VOS, the users only draw scribbles and only a small number of pixels is annotated at each turn. Thus, a generated whole map at one turn contains inadequate information of the whole object, as shown in Figure. 4.5 (a). Therefore, we construct a memory module for storing as well as accumulating the past whole maps to enhance the discriminative representations from target objects. Denote $\mathbf{M}^w \in \mathbb{R}^{m,o,h,w}$ as the whole memory module. The four dimensions of $\mathbf{M}^w$ (m, o, h, w) means total frame number of a video, object number, height and width of the whole matching maps. The maximum of the whole map is 1, which means the distance is largest. Thus, the whole map memory $\mathbf{M}^w$ is initialized with the value of 1 and updated by remaining the minimum value of pixels at each turn. Figure. 4.5 (a) shows the updating procedure of the whole memory. At the turn t, for the frame of i, the *writing* process of memory $\mathbf{M}^w$ is

$$\mathbf{M}^w_{i,t} = \min(\mathbf{M}^w_{i,t-1}, \mathbf{W}_{i,t}). \tag{4.3}$$

The *reading* process of the whole memory is simple - we directly use the aggregated whole map, $\mathbf{M}^w_{i,t}$.

Memory Module for the Partial Matching Map. The apparent differences between two near frames are limited. Thus, the local part is similar across frames. We transfer the predicted segment map from the $(i-1)^{th}$ to i^{th} frame in a local region for each pixels. Similar to [92], we propose the partial matching map. Denote $\mathbb{P}_{i-1,o}$ as the pixels assigned to the object o at the $(i-1)^{th}$ frame. The neighborhood set $\mathbb{B}(p)$ is a set consisting of the pixels at most k pixels away from p. Right of Fig. 4.3 shows how partial maps are generated.

For a pixel p from the i^{th} frame, the partial map $\mathbf{P}_{i,t}$ can be calculated utilizing

$$\mathbf{P}_{i,t}(p) = \begin{cases} \min_{q \in \mathbb{P}^N_{i-1,o}} dist(p,q) & \text{if } \mathbb{P}^N_{i-1,o} \neq \varnothing \\ 1 & \text{else.} \end{cases} \tag{4.4}$$

$\mathbb{P}^N_{i-1,o} \coloneqq \mathbb{P}_{i-1,o} \cap \mathbb{B}(p)$ denotes the intersection of the set of pixels from the $(i-1)^{th}$ frame $\mathbb{P}_{i-1,o}$ and the neighbourhood set $\mathbb{B}(p)$ around p.

For the interactive frame, the pixels labeled by user-scribbles are reliable. Differently, the predictive segment masks of the $(i-1)^{th}$ frame are unreliable because of the incorrect predictions. Besides, the predicted errors will aggregate during the transference operation. Thus, if the present i^{th} frame is temporally distant from the $\hat{i}^{th}$ frame,

the performance of the predicted segmentation mask drops a lot. Thus, we construct a memory module $\mathbf{M}^p \in \mathbb{R}^{m,t,o,h,w}$ for storing partial maps. m, t, o, h, w denote the frame number, the turn number, the object number, the height and width of partial maps. We employ $\mathbf{M}^p \in \mathbb{R}^{m,t,o,h,w}$ to directly store the past partial maps from early interactive turns. The *writting* process of the partial map memory module is

$$\mathbf{M}^p_{i,t} = \mathbf{P}_{i,t}. \tag{4.5}$$

$\mathbf{P}_{i,t}$ denotes the partial map of the i^{th} frame.

The local memory reading is processed by directly using the temporally nearest interactive frame when processing the i^{th} frame. Concretely, When processing the i^{th} frame, we compute the temporal distance between the present i^{th} frame and the user-annotated $\hat{i}^{th}$ frame of every turn t, $dist_t = |i - \hat{i}_t|$. We use the closest map to the user-annotated $\hat{i}^{th}$ frame as the partial map. Nevertheless, we observe that segmentation accuracy at later turns performs better than the former turns. Therefore, the present frame utilizing the partial map at turn 8 performs better than using the partial map at turn 1, although it is temporally distant at turn 8. Thus, we leverage the closest partial map from the interactive frame at previous T turns. Partial maps of beginning turns will be dropped. $T = 1$ indicates we utilize partial matching maps of the present turn while use no memory module. Formally, denote the final partial map of frame i at turn t^* as $\mathbf{P}'_{i,t^*}$. The *reading* process of $\mathbf{P}'_{i,t^*}$ from $\mathbf{M}^p$ is

$$\mathbf{P}'_{i,t^*} = \mathbf{M}^p_{i,t'}, t' = \arg\min_t |i - \hat{i}_t| \ and \ |t' - t^*| \leq T \tag{4.6}$$

The memory mechanism for the partial map is shown in Fig. 4.5 (b).

We propose a transference head composed of four Conv layers to generate the logits for one object. The input of the transference head is the ensemble of the pixel vector map, the whole and partial map from memory modules, and the segment mask of the $(i-1)^{th}$ frame. The logits are concatenated and fed into a softmax layer to produce the probability map.

Interaction Branch. This interaction branch predicts the segmentation mask for the user-annotated $\hat{i}$ frame using scribbles drawn by users. As shown in Fig. 4.2, we assemble the pixel vector map, user-drawn scribbles and the produced segment result of previous turn to generate the segment result of the $\hat{i}$ frame. A "shallow" segmentation head is adopted for the prediction.

In turn-based interactive VOS, the interactive frame needs to record the discriminative information of user annotations, to enhance the predicted mask of the identical

frame at the following turns. However, the annotated scribbles only contain incomplete information of the missing parts, which may still lead to unsatisfactory predictions for some challenging regions. To tackle this issue, an enhanced matching map is proposed to enrich the information of the incomplete scribbles. Because of the similarity between the annotated pixels within the scribbles and the nearby ones of the identical object in the pixel vector space, the enhanced matching map is generated and put into the whole map memory $\mathbf{M}^w$. Denote $\mathbb{P}_{\hat{i}}$ as the set of all pixels of the interactive $\hat{i}$ frame and $\mathbb{P}_{\hat{i},o}$ is the set of user-labeled pixels of the object o. The distance of a pixel p's nearest neighbor in $\mathbb{P}_{\hat{i},o}$ is calculated to generate the enhanced map. Although we assume that pixel embeddings of an identical object should be near to each other, the pixels in a local region, who represent a similar appearance, are closer than pixels of the same object far away in the spatial domain. Thus, for a pixel $p \in \mathbb{P}_{\hat{i}}$, we utilize pixels in its partial neighbor for computing the distance. q is denoted as any pixel in a neighbour set $\mathbb{B}(p)$ of p. $\mathbb{B}(p)$ has pixels at most k pixels away from p. Thus, we can obtain the enhanced map $\mathbf{E}_{\hat{i}}(p)$ with respect to p by

$$\mathbf{E}_{\hat{i}}(p) = \begin{cases} min_{q \in \mathbb{P}^N_{\hat{i},o}} dist(p,q) & \text{if } \mathbb{P}^N_{\hat{i},o} \neq \varnothing \\ 1 & \text{else.} \end{cases} \tag{4.7}$$

$\mathbb{P}^N_{\hat{i},o} \coloneqq \mathbb{P}_{\hat{i},o} \cap \mathbb{B}(p)$ denotes the intersection of the user-labeled set $\mathbb{P}_{\hat{i},o}$ and the neighbour set $\mathbb{B}(p)$. As shown in Fig. 4.4, the enhanced map has more informative knowledge of the targets than scribbles. The enhanced map $\mathbf{E}_{\hat{i}}$ is stored and accumulated in the whole memory $\mathbf{M}^w$. $\mathbf{M}^w$ is written by

$$\mathbf{M}^w_{\hat{i},t} = \min(\mathbf{M}^w_{\hat{i},t-1}, \mathbf{E}_{\hat{i},t}). \tag{4.8}$$

This memory module is able to boost the segment performance of the interactive frame at following turns.

4.3 Experiments

4.3.1 Training

The training procedure of our method includes two periods. First, our model is trained with the backbone and the transference branch. The transference procedure takes as input three RGB images: the interactive, the previous and the present frames. We choose three different frames in a video clip to simulate the transferring procedure

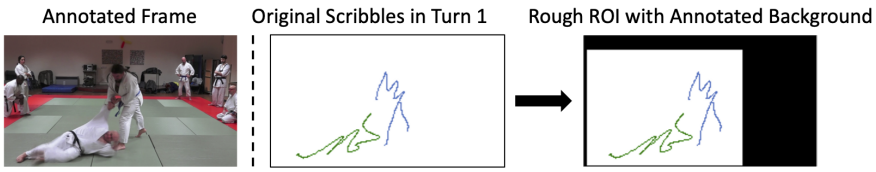

Fig. 4.6 At the first turn the scribbles are drawn only on the fore while no scribbles are on the back. To make our inference consistent for the beginning turn and the following turns, we employ a coarse region of interest (ROI) by assigning the pixels out of the region as the background.

for training. One frame is treated as the interactive frame with user annotations. Another two frames are neighbouring, treated as the previous and the present frame, respectively. Previous approaches [33, 70] choose to simulate user-annotations for the interactive frame when training the transference network. These simulated scribbles are selected from the ground truth. Therefore, the ground-truth masks are actually utilized with the training iteration growing. Thus, we directly employ the ground-truth mask for the interactive frame, since the simulation of scribbles is inefficient during training. We compared the two training strategies and found the performance is similar in practice.

Second, we fixed the trained pixel embedding backbone and the transference segmentation head to train the interaction branch. Collecting a large amount of manually drawn scribbles by users is time-consuming and expensive. Thus, we simulate scribbles during training the interaction branch. At the beginning turn, we employ the user scribbles from the train set supplied by the DAVIS dataset [7]. At the next turns, we simulate scribbles by comparing the predictive masks and the ground truths. At the first turn the scribbles are drawn only in the foreground while no scribbles are in the background. At the following turns, both false positive and true negative scribbles are provided. To address this issue, we utilize the background label as the previous turn's segment results for the first turn.

4.3.2 Inference

Following the turn based interactive VOS proposed by DAVIS [7], users only draw scribbles on the objects while no scribbles on the backgrounds at the beginning turn. To make our inference consistent for the first turn and the succeeding turns, we employ a coarse region of interest (ROI) and assign the pixels out of the region as background, as shown in Fig. 4.6. The coarse ROI involves total positive annotations and is enlarged to adequate size to involve all of the objects. Firstly, the pixel vector maps of entire

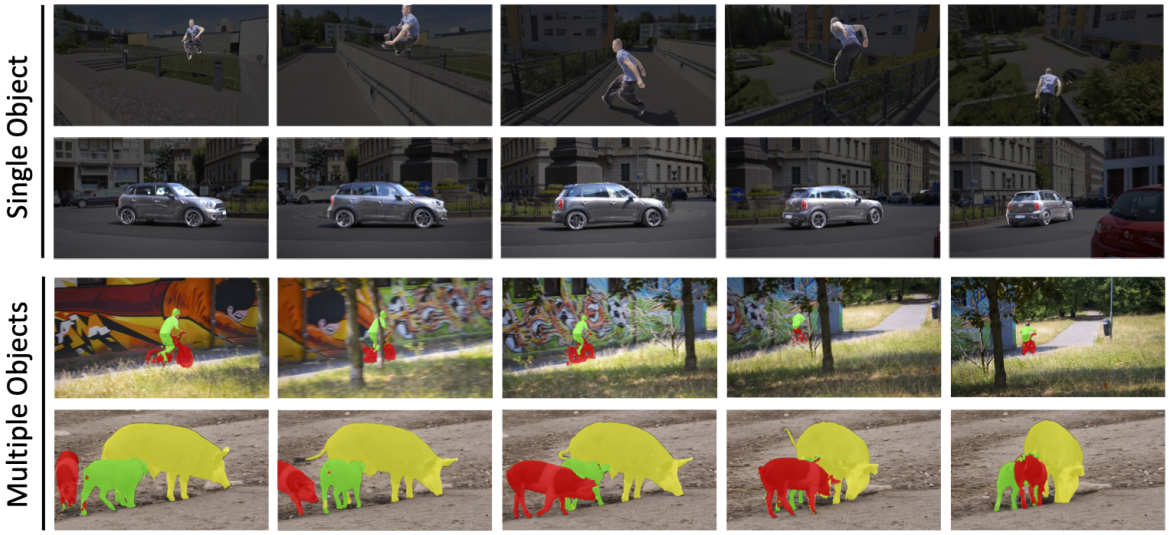

Fig. 4.7 The segment masks on DAVIS are visualized. Users' scribbles are synthesized by the computer, suggested by [7]. Segmentation masks are selected after 8 turns.

frames are produced by the pixel vector encoder. Then the interaction and transference branches are used to predict the segmentation results iteratively. For turns after the first turn, the frame with the worst predicted mask is utilized as the new interactive frame. The pixel vectors are only extracted once at the beginning turn. At the following turns, the pixel vectors are used directly by the segmentation heads in the interaction and transference branch, making our model more efficient than previous approaches.

Implementation and Hyperparameters. We utilize DeepLabv3plus [13] with ResNet101 [28] as the embedding encoder backbone, and the stride of the output feature is 4. We downsample the dimension of the output to 100 utilizing one convolutional layer, whose kernel size is 3×3. For both the interaction and transference branches, we utilize a segmentation head respectively with four depth-wise separable convolutional layers. The kernel size of these layers are 7×7, and the batch normalization and the ReLU activation is adopted after every convolutional layer. The middle dimension of the convolutional layer is 256. After that we use a 1×1 convolutional layer to predict logits.

For the partial map generation, the pixel vector map's dimension is reduced by a factor of 2. The local window size is set to $k = 12$, balancing the trade-off between efficiency and accuracy. A SGD optimizer is utilized for optimization. We use the adaptive bootstrapped loss [74], utilizing 100% to 15% hardest pixels from step 0 to step 50000. The learning rate is initialized by 0.0007 and the batch size is set to 2. The data augmentation strategies, including random scaling, random flipping and random

cropping, are used for the input images. The size of the input image is 416×416. For the first training stage, the weights of the DeepLabv3plus backbone are pre-trained on COCO [56]. The pixel embedding backbone and the transference head are trained on DAVIS [75] for 100000 steps. For the second training period, we employ a turn-based training procedure with three turns per loop. Following DAVIS Challenge 2018, we use only the foreground scribbles at the first turn and both the foreground and background scribbles at the following turns. We train the second period on DAVIS [75] for 80000 steps.

4.3.3 Segmentation Results

It is difficult to evaluate the interactive VOS quantitatively because different users draw different scribbles. Thus, DAVIS Challenge *et al.* [7] utilizes the computer to simulate user interactive actions.

Qualitative Performance. In Fig. 4.7, the segmentation results on the validation set of DAVIS2017 are visualized qualitatively. We observe that the proposed model has the ability of generating precise segmentation masks on both the single and multiple objects conditions. The 3rd row of Fig. 4.7 shows that our model is able to tackle the occluded situations. The 4th row of Fig. 4.7 illustrates that when meeting hard examples, for instance, there are several objects of an identical category which need to be segmented, and these objects are occluded mutually, our method may predict error results in similar regions of different objects. One possible reason is that the pixel embeddings of similar regions are close in the embeddings space.

Quantitative Performance. We evaluate the proposed approach on DAVIS [7], utilizing the interactive track. In this track, the computer is used to draw scribbles on the predicted results for 8 turns. The evaluation metrics include the Jaccard at 60 seconds (J@60s) and the area under the curve (AUC). Comparison results between our approach and previous state-of-the-arts are shown in Table. 4.1. Our model surpasses all previous approaches. Compared with the best competing method Heo [33] based on the accuracy, our approach outperforms it by around +4.7% AUC. Comparing with the best competing method Oh *et al.* [70] based on inference time, the proposed approach outperforms it by around +2.7% J@60s. Different from previous methods, our method use no bells and whistles, *i.e.*, the CRF post-processing, the optical flow, or an extra train set YoutubeVOS [99]. Besides, our method can complete 7-turn

Method	+OF	+CRF	+YV	AUC	J@60
Najafi *et al.* [67]		✓		70.2%	54.8%
Heo *et al.* [33]			✓	69.8%	69.1%
Heo *et al.* [33]	✓		✓	70.4%	72.5%
Oh *et al.* [70]			✓	69.1%	73.4%
MA-Net(Ours)				**74.9%**	**76.1%**

Table 4.1 Performances of the proposed method and previous approaches on DAVIS validation set. The performances are sorted based on J@60. We denote +OF, +CRF, +YV as adding the optical flow, the CRF [47] post-processing and the extra YoutubeVOS train set [99].

Partial window size k	6	9	12	15
AUC	72.4	73.7	74.9	74.8
J@60	73.0	75.3	76.1	76.1

Table 4.2 The ablations on partial window size k.

interactions within 60 seconds while previous state-of-the-art [70] can complete 5-turn only, indicating the efficiency of our method. [1]

4.3.4 Ablation Studies

Impact of Memory Modules. Ablation studies about the proposed two memory modules, the whole and the partial memories, are conducted on the DAVIS2017 dataset [75]. Fig. 4.8 and Fig. 4.9 show the ablation studies based on the J@60 score when increasing number of turns. Fig. 4.8 shows the methods with/without the whole and the partial memory modules. **No Whole** denotes our method without the whole memory. Thus only the whole maps at the first turn are employed for mask generation. **No Partial** denotes our method without the partial memory. The partial matching map at the current turn is used for mask generation. **No Whole and Partial** denote our model without the whole and partial memory modules. Fig. 4.8 illustrates that both our proposed partial and whole memory modules take effects for the interactive VOS and boost the performance significantly because of leveraging all user-annotation information from previous multiple turns.

For the partial map memory module, the partial map nearest to the interactive frame is utilized in past T turns. There exist trade-off situations between selecting the closest frame and the nearest turn. The segmentation results using the partial

[1] To fairly compare the efficiency, we evaluate our method on a 1080Ti GPU, which is the same as Oh [70].

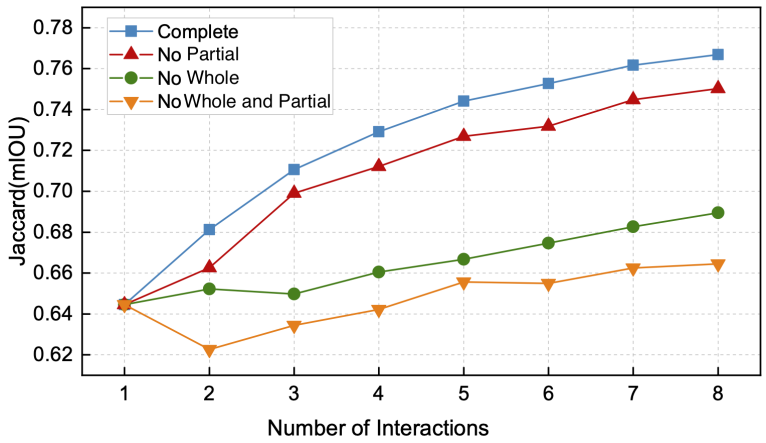

Fig. 4.8 The impact of the proposed memory modules. All experiments are conducted on DAVIS.

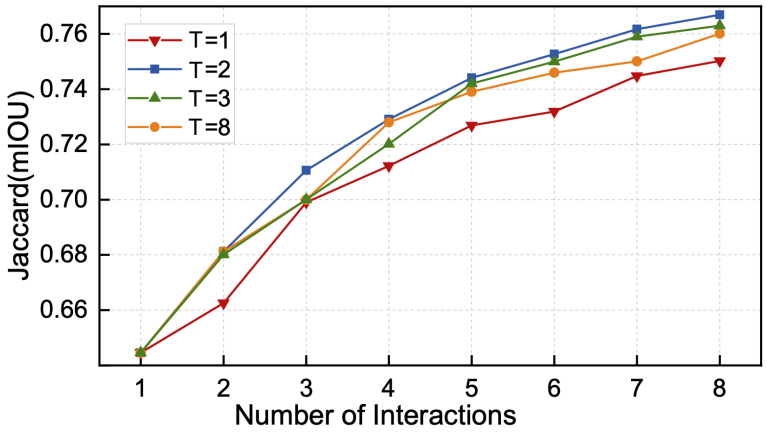

Fig. 4.9 Ablation studies on T in the partial map memory module. T indicates that partial maps in past T turns in the memory module are utilized.

map distant from the interactive frame performs worse because of accumulating the errors during the transference. Thus we select the partial map at the turn where it is closest to the interactive frame. However, the segmentation accuracy becomes better and better with the interaction turns growing. Thus we select the closest partial map to the interactive frame in past T turns. $T = 1$ indicates we utilize the partial map at the present turn while $T = 8$ indicates we utilize the partial map in all past turns. As shown in Fig. 4.9, we found that the performance of $R > 1$ is better than $R = 1$, illustrating the effectiveness of the partial memory module. We choose $T = 2$ where the proposed model obtains the best segmentation results.

Impact of the Enhanced Map. For the interaction branch, we propose an enhanced map to enhance the scribbles and store it in the whole memory module. The

enhanced map stored in the memory can provide useful information for the interactive frame in next turns. In addition, since the predicted mask of the interactive frame will affect other frames by transference, thus, the improvements of the user-annotated frames brought by the enhanced map will bring additional benefits to the other frames during transference. We did ablation studies about the enhanced map. If we do not store the enhanced map in the whole memory, the AUC score will decrease from 0.749 to 0.744.

Ablations on the Partial Window Size. Ablation studies about the local window size k are conducted. Smaller k makes the inference more efficient because of fewer computations. Nevertheless, a small k will influence the performance of the method. As shown in Table. 4.2, we select $k = 12$ for best results.

4.4 Conclusion

Interactive VOS aims to segment objects in an entire video clip when the user annotations are provided, for instance, scribbles. We propose a unified and efficient framework to address the interactive VOS. Our model is composed of three parts, the pixel vector encoder, the interaction and the transference branches. These interaction and transference branches share the pixel vector encoder, making our model more efficient than previous approaches. Besides, we employ two memory modules to store and accumulate the informative knowledge of users' annotation and predictive results at previous turns, boosting the segmentation performance. A future work of our method is improving the interaction branch by further augmenting the user annotations (*e.g.*, scribbles). For instance, the scribbles and the predicted edge information can provide more sufficient informative knowledge to generate masks with higher quality.

Chapter 5

A Generic Bayesian Framework for Few-shot Image Classification

5.1 Introduction

The chapter presents a generic Bayesian framework for the few-shot image classification. The purpose of few-shot classification is adapting new tasks after seeing a few examples. For the few-shot image classification, models are learned with the support set and evaluated with the query set for one task. As all we know, the deep learning models are eager for massive labeled data. However, the few-shot image classification is short of data with labels and introduces nondeterminacy. Thus, a robust deep model for the few-shot classification needs the ability of predicting accurate results, as well as reasoning about the uncertainty. Concretely, the deep model is able to predict the confidence score of the results accurately - a lower confident score should be predicted if the classification result is unreliable. Good uncertainty estimation can also be applied to distinguish out-of-distribution examples and thus helps to avoid making incorrect predictions. Besides, some few-shot classification approaches endure the overfitting problem [83, 91] during training because of the lack of data.

Bayesian methods have the ability to address the mentioned problems and estimate the uncertainty [11, 59, 73, 82]. Previous Bayesian methods [20, 78, 104] are usually based on an optimization-based framework, model-agnostic meta-learning (MAML) [20]. For each task, these methods need to adapt the meta parameters to the posterior distributions over the task-specific parameters by optimization. The optimization with back-propagation during the inference is time-consuming and needs more GPU-memory.

Metric-based few-shot learning methods [83, 89, 91] aim at learning an embedding space in which representations of the identical class are close and vice versa. These

methods need no additional parameters optimization during inference, which is simpler than optimization-based algorithms. In our proposed framework, a Bayesian model is developed to build the relationship between the query and gallery set utilizing a discriminative function. Thus, there is no need to optimize the parameters when testing.

A key challenge is that the computation of posterior distribution is usually difficult because of the unavailable integral term. Recently, some approximation methods for deep neural networks [4, 43, 60] have been proposed. In this dissertation, we employ recently proposed natural-gradient variational inference (NGVI) [36, 43, 73] to estimate the posterior distribution over neural networks' weights. Natural-gradient update exploits the Riemannian space and reforms the dissimilarity of distributions to make the models converge fast.

Different from early deep models for the few-shot classification, our Bayesian method promotes the ability of estimating the uncertainty. Different from previous Bayesian few-shot classification approaches, our approach is more efficient because of no back-propagation operations during testing. Our proposed Bayesian framework is generic, and most metric-based few-shot classification approaches can be adapted to our framework. In this dissertation, we employ the Prototypical Networks (ProtoNet) [83] as the base model to evaluate our method. However, our model is able to be applied to other metric-based models (e.g. GNN few-shot model [22]).

Extensive experiments are conducted to show the effectiveness of our Bayesian model. On two few-shot classification datasets, *i.e.*, *mini*ImageNet and Few-Shot Cifar-100 (FC100), the results show the superiority of uncertainty estimation over the baseline, ProtoNet, and previous Bayesian few-shot classification approaches. Moreover, our model shows superior robustness to adversarial attacks, and alleviate overfitting when training. For the predictive accuracy our model can achieve similar or better results compared with previous methods.

5.2 Preliminaries

The section introduces the preliminary knowledge about the few-shot image classification and an approximation method for the estimation of the posterior distribution: natural-gradient variational inference (NGVI).

5.2.1 Few-Shot Image Classification

In the few-shot image classification, one task π is composed of a *support* set $\mathcal{S}_\pi = \{(\mathbf{x}_1, z_1), ..., (\mathbf{x}_M, z_M)\}$ of M labeled image and a *query* set $\mathcal{Q}_\pi = \{(\mathbf{x}'_1, z'_1), ..., (\mathbf{x}'_{M'}, z'_{M'})\}$ of M' labeled image. $\mathbf{x}_i, \mathbf{x}'_i \in \mathbb{R}^D$ is a D-dimensional vector representing an image while $z_i, z'_i \in \{1, ..., C\}$ is the corresponding category label. We denote $\mathbf{X}_\pi = (\mathbf{x}_1, ..., \mathbf{x}_M)^T$ as the images of $\mathcal{S}_\pi$ and $\mathbf{z}_\pi = (z_1, ..., z_M)^T$ as the labels of $\mathcal{S}_\pi$. Similarly, we denote $\mathbf{X}'_\pi = (\mathbf{x}'_1, ..., \mathbf{x}'_{M'})^T$ and $\mathbf{z}'_\pi = (z'_1, ..., z'_{M'})^T$ as the images and labels of $\mathcal{Q}_\pi$. The few-shot classification aims to recognize the class of images from the query set $\mathcal{Q}_\pi$ given the labeled support set $\mathcal{S}_\pi$. If the support set $\mathcal{Q}_\pi$ has C classes and each class contains K data points, this setting is called the C-way-K-shot classification tasks.

Few-shot learning models are commonly learned by sampling tasks from a dataset. Each few-shot learning task, consisting of a support set and a query set, is named an "episode". Each episode is sampled randomly from a dataset and used to train the few-shot learning models. We split the dataset of tasks into a training set $\mathcal{T}$, a test set $\mathcal{E}$, and a validation set $\mathcal{V}$.

5.2.2 Natural-Gradient Variational Inference

It is hard to calculate the posterior distribution of weights from a neural network directly. Thus, approximate methods have been proposed to estimate the posterior distribution, including MCMC methods [68] and variational inference [4, 60]. Variational Inference(VI) has advantages of efficiency and simplicity, which is used in this dissertation.

Suppose to utilize the variational distribution $q_{\boldsymbol{\eta}}(\boldsymbol{\theta})$ to approximate the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$. Then the objective is:

$$\mathcal{L}(\boldsymbol{\eta}) = -\mathbb{E}_q[\log p(\mathcal{D}|\boldsymbol{\theta})] + D_{\mathrm{KL}}(q_{\boldsymbol{\eta}}(\boldsymbol{\theta})||p(\boldsymbol{\theta})). \tag{5.1}$$

Denote $\mathcal{D}$ as the observed data points and $\boldsymbol{\eta}$ as the variational parameters of $q_{\boldsymbol{\eta}}(\boldsymbol{\theta})$.

We employ a recent variational inference method, the natural-gradient variational inference (NGVI) [36, 43, 44]. Denote $\boldsymbol{\eta}$ as the natural parameter and $\mathbf{m}$ denotes the expectation parameter and assume the variational distribution $q_{\boldsymbol{\eta}}(\boldsymbol{\theta})$ takes the exponential-family form. With the assumption that the exponential-family is in minimal representation, there is a one-to-one mapping between $\boldsymbol{\eta}$ and $\mathbf{m}$. The relationship between $\boldsymbol{\eta}$ and $\mathbf{m}$ is

$$\mathbf{F}(\boldsymbol{\eta})^{-1}\nabla_{\boldsymbol{\eta}}\mathcal{L} = \nabla_{\mathbf{m}}\mathcal{L}, \tag{5.2}$$

where we denote the Fisher information matrix as $\mathbf{F}(\boldsymbol{\eta})$. According to the mirror descent framework [44], we have the natural-gradient descent in the natural-parameter space:

$$\boldsymbol{\eta}_{t+1} = \boldsymbol{\eta}_t - \beta_t \mathbf{F}(\boldsymbol{\eta})^{-1} \nabla_{\boldsymbol{\eta}} \mathcal{L} = \boldsymbol{\eta}_t - \beta_t \nabla_{\mathbf{m}} \mathcal{L}. \tag{5.3}$$

β_t is the learning rate in iteration t.

According to Equation (5.3) we found that the gradient is calculated by the expect parameters $\mathbf{m}$ while the update step is conducted in the natural-parameter space. The natural-gradient VI exploits the Riemannian space of $q(\boldsymbol{\theta})$, which is more advanced than $\mathbf{m}$ in the Eulerian space because of correcting the measure of the dissimilarity of distributions.

Suppose that $q(\boldsymbol{\theta})$ is a Gaussian distribution, with hyperparameters $\boldsymbol{\mu}$ (mean) and $\boldsymbol{\Sigma}$ (convariance). We utilize the mean-field approximation for the Gaussian distribution, which means the covariance matrix is a diagonal matrix: $\boldsymbol{\Sigma} = diag(\boldsymbol{\sigma}^2)$.

The natural parameters $\boldsymbol{\eta}$ and the expectation parameters $\mathbf{m}$ of $q(\boldsymbol{\theta})$ can be written as:

$$\boldsymbol{\eta}^{(1)} := \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \boldsymbol{\eta}^{(2)} := -\frac{1}{2}\boldsymbol{\Sigma}^{-1} \tag{5.4}$$

$$\mathbf{m}^{(1)} := \mathbb{E}_q[\boldsymbol{\theta}] = \boldsymbol{\mu}, \mathbf{M}^{(2)} := \mathbb{E}_q[\boldsymbol{\theta}\boldsymbol{\theta}^T] = \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma}. \tag{5.5}$$

Then Equation (5.3) can be written as:

$$\begin{aligned} \boldsymbol{\sigma}_{t+1}^{-2} &= \boldsymbol{\sigma}_t^{-2} + 2\beta_t[\nabla_{\boldsymbol{\sigma}^2}\mathcal{L}_t], \\ \boldsymbol{\mu}_{t+1} &= \boldsymbol{\mu}_t - \beta_t \boldsymbol{\sigma}_{t+1}^2[\nabla_{\boldsymbol{\mu}}\mathcal{L}_t]. \end{aligned} \tag{5.6}$$

Note that the derivative of $\boldsymbol{\sigma}^2$ needs to calculate the Hessian matrix of the objective, which is complicated. Thus, we utilize the Variational Online Gauss-Newton (VOGN) method [43] using the Generalized Gauss-Newton approximation, which approximates second-order gradient with the square of the first-order gradient.

5.3 Metric-Based Bayesian Framework

We propose a metric-based Bayesian framework to directly model the support and query set without complicated back-propagation for each task. Therefore, we consider a discriminative model which predicts a probability distribution $p(\mathcal{Q}_\pi|\mathcal{S}_\pi, \boldsymbol{\theta})$ of the the query set $\mathcal{Q}_\pi$ given the support set $\mathcal{S}_\pi$ and the model parameter $\boldsymbol{\theta}$. Denote π as one task. We aim to acquire the posterior distribution of model parameters $p(\boldsymbol{\theta}|\mathcal{D})$ where

$\mathcal{D} = \{(\mathcal{S}_\pi, \mathcal{Q}_\pi) | \pi \in \mathcal{T}\}$ and $\mathcal{T}$ is the training set of tasks. After obtaining $p(\boldsymbol{\theta}|\mathcal{D})$, we can predict the query set label by $p(\mathcal{Q}_\pi|\mathcal{S}_\pi, \mathcal{D}) = \int p(\mathcal{Q}_\pi|\mathcal{S}_\pi, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) \mathrm{d}\boldsymbol{\theta}$. According to the Bayes theorem, the posterior distribution of the model parameters is

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto \prod_{\pi \in \mathcal{T}} p(\mathcal{Q}_\pi|\mathcal{S}_\pi, \boldsymbol{\theta}) p(\boldsymbol{\theta}). \tag{5.7}$$

We denote $p(\boldsymbol{\theta})$ as the prior distribution of the model parameters. $p(\mathcal{Q}_\pi|\mathcal{S}_\pi, \boldsymbol{\theta}) := p(\mathbf{z}'_\pi|\mathbf{X}'_\pi, \mathbf{X}_\pi, \mathbf{z}_\pi, \boldsymbol{\theta})$ denotes a discriminative model, predicting the results of the query label by the support set and the model parameters. The metric based few-shot learning approaches can serve as the discriminative model, without complete task-specific optimization over the support set $\mathcal{S}_\pi$ during inference.

In this dissertation we adopt a variational distribution $q(\boldsymbol{\theta})$ to approximate the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$ by the natural-gradient VI method.

5.3.1 Objective Function

To approximate the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$ by $q(\boldsymbol{\theta})$, the KL-divergence between the two distributions is minimized. Formally, the objective function is

$$\begin{aligned}
\mathcal{L}(q(\boldsymbol{\theta})) = \; & D_{\mathrm{KL}}(q(\boldsymbol{\theta}) || p(\boldsymbol{\theta}|\mathcal{D})) \\
= \; & D_{\mathrm{KL}}\left(q(\boldsymbol{\theta}) \middle\| \prod_{\pi \in \mathcal{T}} p(\mathcal{Q}_\pi|\mathcal{S}_\pi, \boldsymbol{\theta}) p(\boldsymbol{\theta})\right) + C \\
= \; & \mathbf{E}_{\boldsymbol{\theta} \sim q(\boldsymbol{\theta})} \log \frac{q(\boldsymbol{\theta})}{\prod_{\pi \in \mathcal{T}} p(\mathcal{Q}_\pi|\mathcal{S}_\pi, \boldsymbol{\theta}) p(\boldsymbol{\theta})} + C \\
= \; & -\mathbf{E}_{\boldsymbol{\theta} \sim q(\boldsymbol{\theta})} \sum_{\pi \in \mathcal{T}} \log p(\mathcal{Q}_\pi|\mathcal{S}_\pi, \boldsymbol{\theta}) \\
& + D_{\mathrm{KL}}(q(\boldsymbol{\theta}) || p(\boldsymbol{\theta})) + C.
\end{aligned} \tag{5.8}$$

From Equation (5.8), two terms except for the constant C need to be optimized. One is the log-likelihood function over the query set $\mathcal{Q}_\pi$ while another is the KL term of the prior and the posterior distribution.

For the likelihood term, we adopt metric-based few-shot learning methods [81, 83, 91] to model the likelihood function $p(\mathcal{Q}_\pi|\mathcal{S}_\pi, \boldsymbol{\theta})$, since the metric-based methods directly model the relationship across the support set $\mathcal{S}_\pi$ and the query set $\mathcal{Q}_\pi$. For the KL term, we suppose the prior distribution $p(\boldsymbol{\theta})$ as a Gaussian distribution, with mean $\mathbf{0}$ and the covariance $\mathbf{I}$: $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{0}, \mathbf{I})$. We use the natural-gradient VI method (Equation (5.6)) to optimize the loss function Equation (5.8).

For the training procedure, the parameters are sampled at step t, $\boldsymbol{\theta}_t \sim \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_t, diag(\boldsymbol{\sigma}_t^2))$, utilizing the Reparametrization Trick [46]. The we optimize our model by updating the variational parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ using Equation (5.6). The gradients are calculated using the VOGN [43] method. For the testing procedure, we learn an optimal variational posterior, $\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}^*, diag(\boldsymbol{\sigma}^{2*}))$, and sample parameters $\boldsymbol{\theta}$ from this distribution. We adopt $p(\mathcal{Q}_\pi|\mathcal{S}_\pi, \boldsymbol{\theta})$ to predict the distribution of the query set. The training and testing algorithms are shown in Algorithms 1 and 2.

The reason why we employ the natural gradient VI method is it exploits the Riemannian space of $q(\boldsymbol{\theta})$ and converges faster. To illustrate the effectiveness, we compare our method with another optimization method, Bayes by Backprop (BBB) [5]. Please refer to Section 5.4 for the ablation results.

5.3.2 Model Architecture

Our proposed Bayesian framework is general and any discriminative model that directly takes as input the support set $\mathcal{S}_\pi$ and outputs the query set $\mathcal{Q}_\pi$ can be applied in our framework. Recently, some metric-based few-shot classification models [81, 83] have been proposed and applicable to our Bayesian framework. We adopt ProtoNet [83] to model the likelihood function $p(\mathcal{Q}_\pi|\mathcal{S}_\pi, \boldsymbol{\theta})$ in this dissertation. Extensive experiments on our model with ProtoNet backbone are conducted. To show the generalization of our framework, we employ another metric-based few-shot classification method, the GNN few-shot learning model [81], as the likelihood function.

ProtoNet-based model. ProtoNet generates a prototype $\mathbf{p}_c \in \mathbb{R}^M$ for each class c by an embedding function $h_{\boldsymbol{\theta}}(\mathbf{x}_i)$:$\mathbb{R}^D \to \mathbb{R}^M$, where $\boldsymbol{\theta}$ denotes the model parameters.

The prototype for a class c is computed by averaging the embeddings of the support images:

$$\mathbf{p}_c = \frac{1}{|\mathcal{S}_\pi^{(c)}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}_\pi^{(c)}} h_{\boldsymbol{\theta}}(\mathbf{x}_i), \tag{5.9}$$

where we denote $\mathcal{S}_\pi^{(c)}$ as the support images of class c.

We employ the Euclidean distance function $d(\cdot)$ as the distance metric. ProtoNet produces a distribution over classes for a query point $\mathbf{x}'$ by the softmax over distances to the prototypes in the embedding space:

$$p_{\boldsymbol{\theta}}(y' = c|\mathbf{x}', \mathcal{S}_\pi) = \frac{\exp(-d(h_{\boldsymbol{\theta}}(\mathbf{x}'), \mathbf{p}_c))}{\sum_{c'} \exp(-d(h_{\boldsymbol{\theta}}(\mathbf{x}'), \mathbf{p}_{c'}))}. \tag{5.10}$$

Thus, we can have the likelihood function by estimating $p(\mathcal{Q}_\pi|\mathcal{S}_\pi, \boldsymbol{\theta})$,

$$p(\mathcal{Q}_\pi|\mathcal{S}_\pi, \boldsymbol{\theta}) = \prod_{(\mathbf{x}', y') \sim \mathcal{Q}_\pi} p_{\boldsymbol{\theta}}(y'|\mathbf{x}', \mathcal{S}_\pi). \tag{5.11}$$

GNN-based model. Here we adopt another metric-based few-shot learning model, the GNN few-shot model [81]. Graph neural networks are based on the local operators of a graph $\mathcal{G} = (V, E)$. Takes as input $\mathbf{Z}^k \in \mathbb{R}^{V \times d_k}$ on the vertices of a graph $\mathcal{G}$ at layer k, and adjacency operators $\mathcal{A}$, a GNN layer $Gc(\cdot)$ outputs results for the next layer $\mathbf{Z}^{k+1} \in \mathbb{R}^{V \times d_{k+1}}$ by

$$\mathbf{Z}^{k+1} = \rho(\sum_{\mathbf{B} \in \mathbb{A}} \mathbf{B}\mathbf{Z}^k \boldsymbol{\theta}_{\mathbf{B}}). \tag{5.12}$$

Denote $\boldsymbol{\theta}_{\mathbf{B}} \in \mathbb{R}^{d_k \times d_{k+1}}$ as the optimization parameters and $\rho(\cdot)$ as a point-wise non-linearity. A multilayer perceptron is adopted to learn the adjacency operator $\mathbf{A}^k$,

$$\mathbf{A}^k_{i,j} = MLP_{\boldsymbol{\theta}'}(abs(\mathbf{Z}^k_i - \mathbf{Z}^k_j)), \tag{5.13}$$

and the adjacency familay $\mathbb{A} = \{\mathbf{A}^k, \mathbf{1}\}$.

We build a node of the graph by taking as input the concatenation of the embedding vector of the image $\mathbf{x}_i$ and the one-hot encoding of its label y_i in the support set $\mathcal{S}$,

$$\mathbf{n}^0_i = (\psi(\mathbf{x}_i), h(y_i)), \tag{5.14}$$

where $\psi(\cdot)$ is convolutional neural networks while $h(\cdot)$ is a one-hot encoding operator. For an image $\mathbf{x}'_i$ from $\mathcal{Q}$, the node is the concatenation of the embedding feature and a

uniform distribution. A softmax layer is adopted to the output node of the images in $\mathcal{Q}$, which predicts $p(\mathcal{Q}_\pi|\mathcal{S}_\pi,\boldsymbol{\theta})$. please refer to [81] for more details.

Algorithm 1: Training

Input: Dataset $\mathcal{D}$, Sample number M, Learning rate β
Initialize $\boldsymbol{\mu}_0$, $\boldsymbol{\sigma}_0^2$
$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{0},\mathbf{I})$
while not done **do**
  Sample batch of *episodes* $\{\mathcal{S}_j, \mathcal{Q}_j\} \sim \mathcal{D}$
  for $i = 1$ **to** M **do**
    Sample $\boldsymbol{\theta}_i \sim \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_t, diag(\boldsymbol{\sigma}_t^2))$
    Evaluate $\nabla^i_{\boldsymbol{\mu}_t}\mathcal{L}_t(\mathcal{S}_j,\mathcal{Q}_j,\boldsymbol{\theta}_i)$,$\nabla^i_{\boldsymbol{\sigma}_t^2}\mathcal{L}_t(\mathcal{S}_j,\mathcal{Q}_j,\boldsymbol{\theta}_i)$[1]
  end for
  $\nabla_{\boldsymbol{\mu}_t}\mathcal{L}_t = \frac{1}{M}\sum_{i=1}^{M}\nabla^i_{\boldsymbol{\mu}_t}\mathcal{L}_t$
  $\nabla_{\boldsymbol{\sigma}_t^2}\mathcal{L}_t = \frac{1}{M}\sum_{i=1}^{M}\nabla^i_{\boldsymbol{\sigma}_t^2}\mathcal{L}_t$
  update:
  $\boldsymbol{\sigma}_{t+1}^{-2} = \boldsymbol{\sigma}_t^{-2} + 2\beta_t[\nabla_{\boldsymbol{\sigma}^2}\mathcal{L}_t]$
  $\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t - \beta_t\boldsymbol{\sigma}_{t+1}^2[\nabla_{\boldsymbol{\mu}}\mathcal{L}_t]$
end while

Algorithm 2: Evaluation

Input: Support set $\mathcal{S}_\pi$, Query Set $\mathcal{Q}_\pi$, Sample number M, Learned variational distribution $\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}^*, diag(\boldsymbol{\sigma}^{2*}))$
for $i = 1$ **to** M **do**
  Sample $\boldsymbol{\theta}_i \sim \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}^*, diag(\boldsymbol{\sigma}^{2*}))$
  Compute $p(\mathcal{Q}_\pi|\mathcal{S}_\pi,\boldsymbol{\theta}_i)$
end for
$p(\mathcal{Q}_\pi|\mathcal{S}_\pi) = \frac{1}{M}\sum_{i=1}^{M}p(\mathcal{Q}_\pi|\mathcal{S}_\pi,\boldsymbol{\theta}_i)$

5.4 Experiments

Our metric-based Bayesian framework is evaluated on the few-shot image classification. We employ ProtoNet [83] as our baseline since our framework utilizes ProtoNet as the backbone. Thus, for a fair comparison, we utilize exactly the same architecture as ProtoNet.

[1] $\nabla^i_{\boldsymbol{\mu}_t}\mathcal{L}_t$ and $\nabla^i_{\boldsymbol{\sigma}_t^2}\mathcal{L}_t$ are approximated by first-order and square of first-order derivative with respect to sampled $\boldsymbol{\theta}_i$.

Besides, we compare our method with some other Bayesian few-shot learning methods [20, 25, 78, 104] on both accuracy and uncertainty metrics. The uncertainty metrics include the negative log-likelihood, calibration curves, and entropy of out-of-distribution examples. Moreover, in the experiments, our model shows the robustness against adversarial attacks compared with the baseline, and shows the superiority of preventing overfitting. Finally, our model achieves competitive or better predictive accuracy than previous state-of-the-art approaches.

5.4.1 Datasets

Two datasets are used for the few-shot image classification:

***mini*ImageNet** [91] is derived from ImageNet [16] by Vinyals et al., which is specifically designed for the few-shot classification problem. It consists of 600 84×84 images for 100 classes randomly selected from ImageNet. We split *mini*ImageNet with 64/16/20 classes for train/val/test set following Ravi [79].

FC100 [72] is a few-shot classification dataset derived from CIFAR-100 [48]. It is a challenging dataset because of the low resolution. It consists of 600 32×32 images for 100 classes. We follow one previous paper [72] to split it with 60/20/20 classes for train/val/test set.

5.4.2 Experiments Setting and Implementation Details

We use the 1-shot 5-way, 5-shot 5-way and 5-shot 10-way settings on FC100, while 1-shot 5-way and 5-shot 5-way classification settings on *mini*ImageNet. The baseline, ProtoNet, trains the model by a higher way (30-way episodes during training for 1-shot 5-way classification and 20-way episodes during training for 5-shot 5-way classification). We train our model with the same settings to fairly compare with ProtoNet. We use Adam [45] to optimize ProtoNet following the original paper [83].

We employ the same architecture as our baseline, ProtoNet [83], which consists of four convolutional blocks. Each block is composed of a 64-filter 3×3 convolutional layer, a batch normalization layer [41], a ReLU layer and a 2×2 max-pooling layer. When the input size is 84×84 (*mini*ImageNet), the dimension of the embedding feature is $1,600$, while when the input size is 32×32 (FC100), the output dimension of the embedding feature is 256. It is fair to compare our method with previous Bayesian [20, 25, 78, 104] methods because the four-convolution architecture is the same.

We employ VOGN [43] as the optimization method. We set the learning rate as 0.01, which is decreased by a factor of 10 every $10,000$ iterations. The number

Table 5.1 Negative Log-likelihood (NLL) results on *mini*ImageNet and FC100. Lower is better.

Method	*mini*ImageNet		FC100		
	1-shot,5-way	5-shot,5-way	1-shot,5-way	5-shot,5-way	5-shot,10-way
MAML [19]	1.370±0.022	1.108±0.011	-	-	-
Amortized VI [25]	1.328±0.024	1.165±0.010	-	-	-
VERSA [25]	**1.183**±0.023	0.859±0.015	-	1.253±0.018	1.894±0.009
ProtoNet(baseline) [83]	1.227±0.020	0.854±0.017	1.606±0.016	1.321±0.015	2.013±0.011
BBB[5]	1.232±0.016	0.862±0.015	-	-	
Ours	1.214±0.016	**0.812**±0.013	**1.500**±0.012	**1.199**±0.011	**1.810**±0.008

of total iterations is 30,000 for *mini*ImageNet, while 20,000 for FC100. We sample an episode from the dataset with 15 queries per class for an iteration. During the training process we evaluate our dataset on the validation set and employ the early-stop strategy, which means we select the model that achieves the best performance on the validation set. The hyperparameters of the distribution of model parameters are initialized with mean $\boldsymbol{\mu} = \mathbf{0}$ and precision $\frac{1}{\boldsymbol{\sigma}^2} = \mathbf{10}$. Following previous Bayesian few-shot approaches [25, 78], we sample parameters 10 times from the distribution for training and testing, respectively. To optimize the objective function in Equation (5.8), we need the total number of tasks $\mathcal{T}$. However, for the few-shot classification, tasks are sampled from a train set and the amount is exceedingly large. Considering the limited training iterations, we can only use a subset of $\mathcal{T}$. It is difficult to estimate the subset. So we view the amount of the subset as a hyperparameter. We set the task size 100,000 for *mini*ImageNet and 10,000 for FC100.

5.4.3 Uncertainty Estimation

We propose a probabilistic model for the few-shot classification, which has the ability of reasoning about the uncertainty. In this section, we compare our model with the baseline and previous Bayesian few-shot learning approaches on several metrics, including *Negative Log Likelihood*, *Calibration* and *Entropy of Out-of-distribution Examples*. Then we illustrate the classification accuracy curve of the train and test set when training, demonstrating that our model alleviate the overfitting phenomenon.

Negative Log Likelihood. Log likelihood is a common uncertainty metric. Lower negative log likelihood (NLL) means better uncertainty estimation. We randomly select 600 *episodes*, while each *episode* consists of 15 query samples per class. The total number of test samples is $N_{test} = 600 \times 15$. For a image $\mathbf{x}_i \in \mathcal{Q}$ we denote

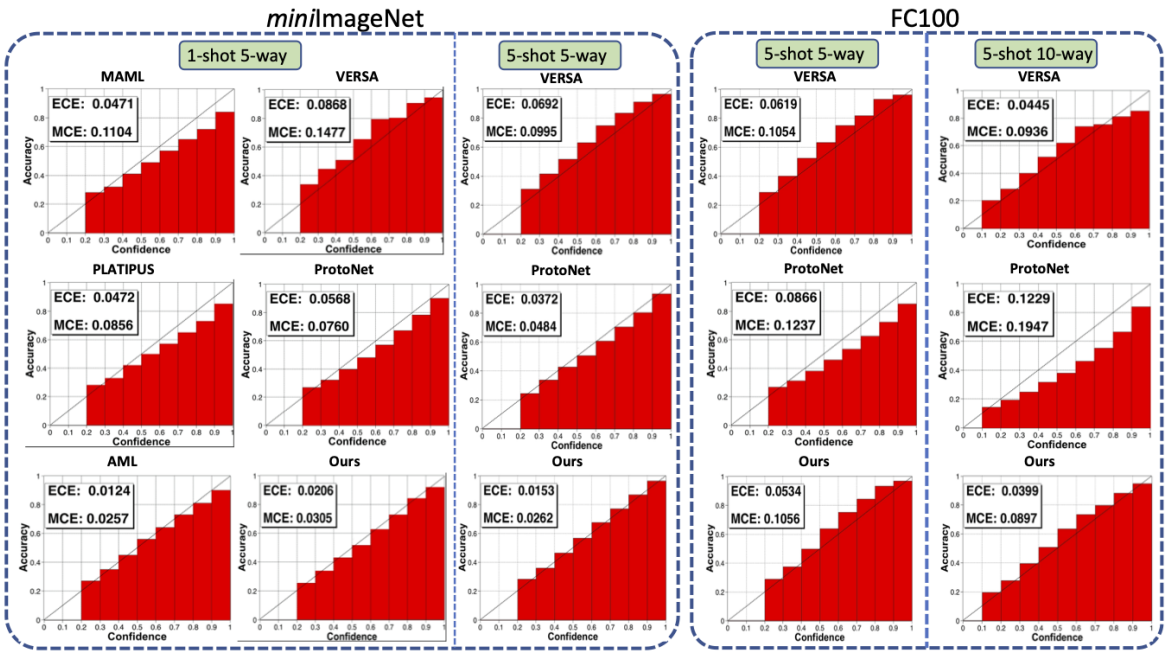

Fig. 5.1 Comparison between our approach and previous state-of-the-arts on the calibration curves and error scores. Bars closer to the diagonal line or lower ECE/MCE means better calibrated. Bars under the diagonal line or over the diagonal line indicates overconfidence or underconfidence, respectively.

a corresponding true label $\mathbf{z}_i$, a 1-of-K encoded vector. We denote the predicted probability of a category c for an image $\mathbf{x}_i$ as p_{ic}, where $c \in 1, ..., C$. The negative log likelihood of test set is $\text{NLL} := -\frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \sum_{c=1}^{C} y_{ic} \log p_{ic}$.

The results on NLL are shown in Table 5.1. We compare our proposed model with MAML [19], the baseline (ProtoNet) [83] and Bayesian methods (Amortized VI and VERSA) [25]. Our method obtains similar or lower NLL than other methods on two few-shot learning datasets, illustrating the superiority of our method over this uncertainty estimation.

Calibration. The ability of reasoning about the uncertainty means the models can obtain informative confidence scores - the predicted probability of one class label corresponds with how likely this prediction is correct. The calibration curves [27] are used to evaluate the calibration of a model. The calibration curves are functions of the expected accuracy given the confidence as input. If the bars of the calibration curves are close to the diagonal curve, the predicted confidence of a class label is close to its ground truth accuracy, which means the model is well-calibrated.

In addition, we also employ two metrics, Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) [66], to assess the calibration. The ECE score is

Table 5.2 Comparison about Adversarial Attacks on *mini*ImageNet.

Method	1-shot,5-way		5-shot,5-way	
	Origin Input	Attacked Input	Origin Input	Attacked Input
MAML [19]	48.7±1.84	1.47±0.19	63.11±0.92	1.80±0.25
VERSA [25]	53.40±1.82	6.05±0.44	67.37±0.86	7.22±0.38
ProtoNet(baseline) [83]	49.42±0.78	1.68 ±0.18	68.20±0.66	1.92±0.20
Ours	50.32±0.82	8.50±0.37	68.64±0.67	9.03±0.38

measured by weighted-averaging the bins' accuracy/confidence difference, while the MCE score means the biggest bins' accuracy/confidence difference. The comparison between our approach and the baseline (ProtoNet) and previous state-of-the-art Bayesian methods (PLATIPUS [20], AML [78] and VERSA [25]) on the calibration is shown in Fig 5.1. The results show that our Bayesian framework performs better on the model calibration, compared with the baseline and previous Bayesian few-shot classification models [20, 25]. Compared with AML [78], although it is slightly better calibrated than ours, its classification accuracy is much lower than ours (45.0% vs. 50.3% for 1-shot 5-way setting).

We observe that for the non-Bayesian method (MAML [19], ProtoNet [83]) the bars are under the diagonal curve, which means these methods tends to predict classification results overconfidently. Bayesian methods have the potential to relieve this overconfidence phenomenon. A competitive Bayesian method, VERSA, predicts the results underconfidently (bars over the diagonal curve). Our approach obtains a balance that makes our model well-calibrated.

Entropy of Out-of-distribution Images. A robust model with a good ability to estimate uncertainty can tackle the examples of unseen classes properly. A good model should be uncertain about unseen images. For example, given a 10-way classifier, an example sampled from unseen classes should be predicted to each class with confidence of $\frac{1}{10}$, which means the classifier does not know how to identify the out-of-distribution image.

We propose the entropies on the out-of-distribution examples to evaluate the quality of uncertainty. Higher entropies mean better uncertainty. The uniform distribution indicates the highest entropy for the classification task. Similar to one previous paper [60] we employ the empirical CDF of entropies on out-of-distribution images to evaluate the performance. The CDF curve towards the bottom-right is better because the rate of the high confidence result on out-of-distribution images is low. 2000 out-of-

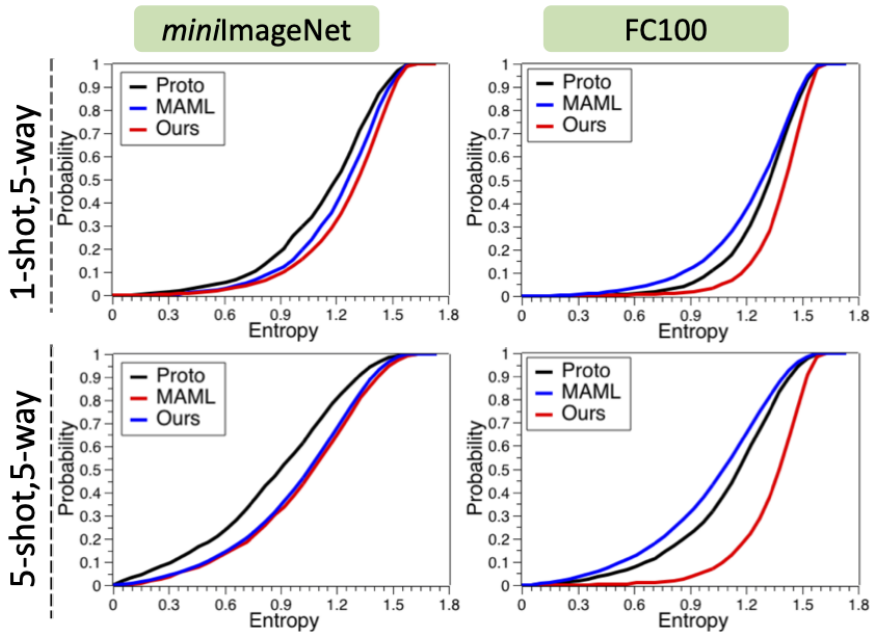

Fig. 5.2 The empirical CDF of entropies for the out-of-distribution examples. Ccloser to right-and-bottom means better uncertainty estimation.

distribution images are selected from the test set. For an out-of-distribution image $\mathbf{x}_i$, we denote the probability of class c for $\mathbf{x}_i$ is p_{ic}, where $c \in 1, ..., C$. The entropy is computed by $H(\mathbf{x}_i) = -\sum_{c=1}^{C} p_{ic} \log p_{ic}$. Fig 5.2 shows the comparison between our Bayesian method and ProtoNet & MAML. Our method predicts better uncertainty results on out-of-distribution images.

5.4.4 Comparison about Adversarial Attacks

Adversarial attacks add small perturbations to input images and significantly affect the performance of neural networks. We conduct experiments to show the robustness of our method against adversarial attacks. We employ the FGSM [24] method to attack our Bayesian model and competitive methods (the baseline, MAML and VERSA). 600 episodes from *mini*ImageNet are sampled and the perturbations are added to the sampled images. The magnitude of the attack perturbation is 0.05. Table 5.2 shows that our model is more robust to attacks compared with the competitive methods. Specifically, our model achieves +7% higher than the baseline (ProtoNet), indicating that the Bayesian model is more robust than the non-Bayesian models.

5.4.5 Comparison about Overfitting

We compare the classification accuracy curve of the train and test set when training between our approach and the baseline. In Fig 5.3, the black and red line denotes the classification accuracy of the train and test set, respectively. We found that the baseline method trained by Adam [45] is easy to overfit to the train set, because the accuracy of the train set improves fast well the accuracy of the test set drops quickly. Differently, our Bayesian method prevents overfitting, since the accuracy of the test set becomes higher with the accuracy of the train set growing.

5.4.6 Image Classification Accuracy Results

Tables 5.3 shows the classification accuracies on two few-shot learning datasets, *mini*ImageNet and FC100. Following ProtoNet, we sample 600 *episodes* to evaluate the test set and for each *episode*, every class contains 15 query images.

In Table 5.3, the first/second/third/fourth group contains non-Bayesian approaches, other Bayesian methods, our baseline, ProtoNet, and our method (BBB means Equation (5.8) optimized by Bayes-by-Backprop), respectively. In the left of Table 5.3, our model achieves competitive or better classification accuracy than other Bayesian methods on *mini*ImageNet, which indicates that our model learns a deep network predicting more robust prototypes.

The right of Table 5.3 shows the comparison between our approach with the ProtoNet baseline and a Bayesian few-shot learning approach, VERSA [25] on FC100. Results show that on 1-shot 5-way, 5-shot 5-way and 5-shot 10-way settings, our model surpasses the ProtoNet baseline and VERSA significantly.

5.4.7 Ablation Study

Comparison with Bayes-by-Backprop. NGVI [44] exploits the Riemannian space of $q(\boldsymbol{\theta})$, corrects the measure of the dissimilarity of distributions and has better convergence. We conduct experiments on NGVI and a previous optimizer for the Bayesian networks, the Bayes-by-Backprop (BBB) [5]. Different from NGVI, the BBB method exploits the Euclidean space. Experiments are conducted on *mini*ImageNet utilizing 1-shot, 5-way, and 5-shot, 5-way settings. BBB and NGVI use the same training and testing strategies for a fair comparison.

For the uncertainty estimation, in Table 5.3, NGVI achieves a better NLL score than BBB. Fig 5.4 shows that the NGVI optimization makes the Bayesian model better calibrated. The model optimized by BBB tends to predict under-confidence results

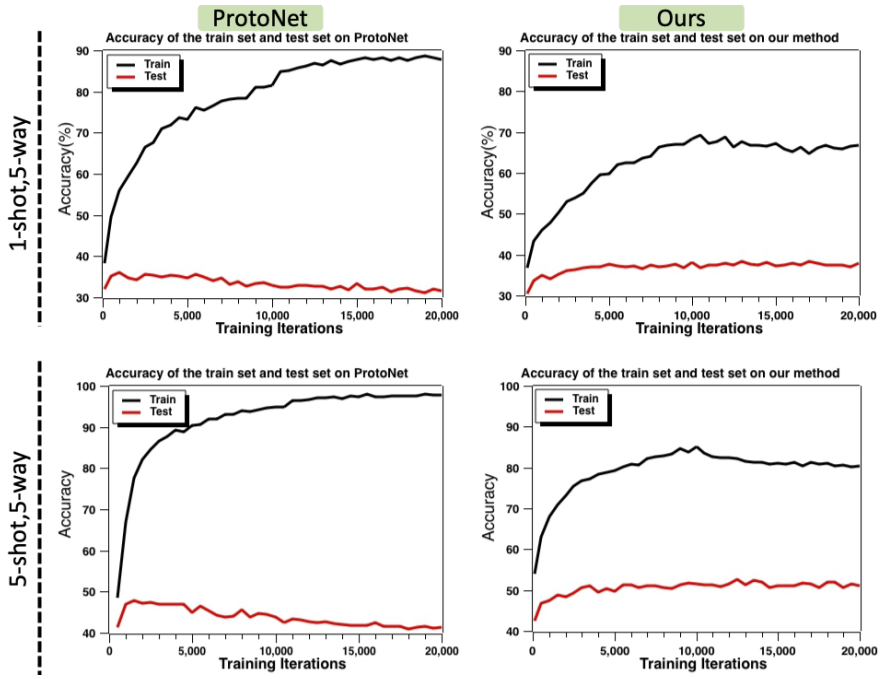

Fig. 5.3 Comparison about the overfitting phenomena on FC100. The black and red lines denote the accuracy of the train and test set, respectively.

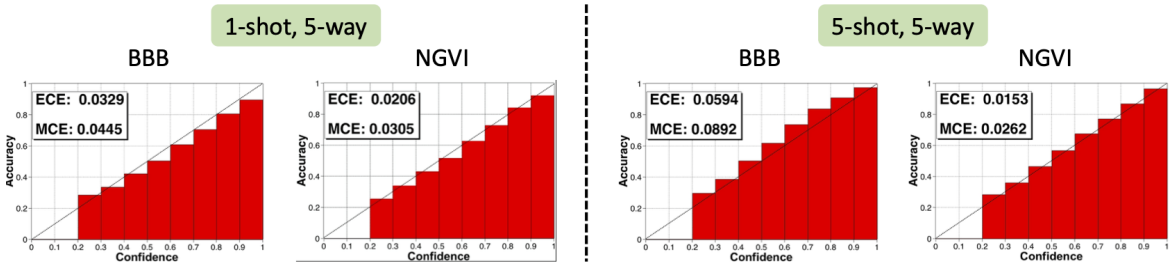

Fig. 5.4 Comparison on the model calibration between BBB [5] and NGVI [36].

(bars over the diagonal curve). For the classification accuracy, the model optimized by NGVI achieves better performance than BBB, as shown in Table 5.3.

Impact of the Sampling Number. In the experiments, we sample the parameters from the distribution $q(\boldsymbol{\theta})$ 10 times for both training and testing processes, which is the same as the competing Bayesian methods [25, 78]. Here we conduct ablation studies on the sampling number when testing. We sample the parameters $1, 3, 5, 10, 15, 20$ times, and show the impact of the sampling number. Fig. 5.5 (1st row) illustrates that the test sampling number M affects the accuracy. With the sampling number growing, the classification accuracy increases. The accuracy increases faster when $M < 10$. When $M \geq 10$ the performances are similar. Fig. 5.5 (2nd row) demonstrates that the sampling number M affects the calibration score. The 1-shot, 5-way setting

Table 5.3 Comparison of accuracy on the *mini*ImageNet dataset and the FC100 dataset. Group 1, 2, 3, 4 denote non-Bayesian approaches, Bayesian approaches, the baseline of ProtoNet and our approach.

Method	*mini*ImageNet		FC100*		
	1-shot,5-way	5-shot,5-way	1-shot,5-way	5-shot,5-way	5-shot,10-way
Matching Nets [91]	46.6	60.0	-	-	-
MAML [19]	48.7±1.84	63.11±0.92	-	-	-
Meta LSTM [79]	43.44±0.77	60.60±0.71	-	-	-
PLATIPUS [20]	50.13±1.86	-	-	-	-
AML [78]	46.00±0.60	-	-	-	-
VERSA [25]	53.40±1.82	67.37±0.86	36.76±1.82	49.46±0.88	33.06±0.58
ProtoNet(baseline) [83]	49.42±0.78	68.20±0.66	36.36±0.62	48.58±0.65	34.02±0.42
BBB[5]	49.35±0.74	67.29±0.66	-	-	-
Ours	50.32±0.82	**68.64**±0.67	**37.79**±0.67	**52.24**±0.69	**35.63**±0.41

and 5-shot, 5-way setting show different patterns. On the 1-shot, 5-way setting, larger M means better calibration. Differently, on the 5-shot, 5-way setting, with M growing, the ECE score shows a "V-shaped" pattern. One possible reason is that when M is small ($M = 1, 3, 5$), the predicted results are overconfident since the accuracy is two low when only providing one example per class (1-shot) while the confidence is high.

Another Model Architecture (GNN). We change the backbone of our Bayesian framework from ProtoNet to GNN, and the experiments are conducted on *mini*ImageNet under the 1-shot, 5-way setting. The training and testing strategies for our model are following the GNN few-shot model [81]. We compare the uncertainty estimation and the classification accuracy between our model and GNN few-shot model. The details about the training and testing strategies are shown in [81].

Negative Log Likelihood (NLL). NLL is an uncertainty metrics and lower NLL means better uncertainty estimation. For the 1-shot, 5-way setting on *mini*ImageNet, our model achieves 1.214 ± 0.017 NLL score while GNN achieves 1.334 ± 0.020. Our model outperforms GNN by 0.12. This demonstrates that our model is effective in reasoning about uncertainty using another backbone model.

Calibration. Fig. 5.6 (a) compares our method with the GNN method on the model calibration. The results show that our model is better calibrated than GNN, and ECE of our model is 0.0789 lower than GNN. Besides, Fig. 5.6 (a) also illustrates that the predictions of GNN are overconfidence while our model depresses this phenomenon.

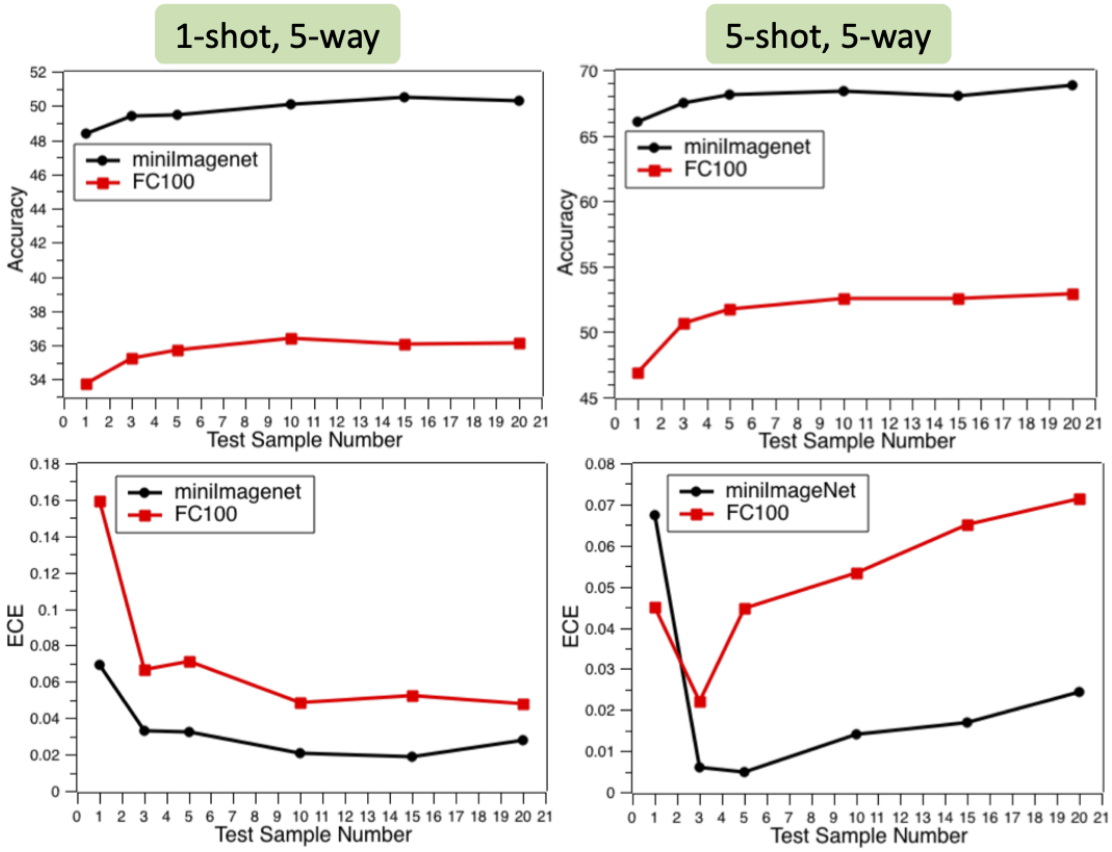

Fig. 5.5 Impact of the Sampling Number on the accuracy (1st row) and the ECE score (2nd row).

Entropy of Out-of-distribution Images. We conduct ablations on the out-of-distribution images. We randomly sample 2,000 out-of-distribution images from the test set in *mini*ImageNet, and compute the entropies of the images. Fig. 5.6 (b) illustrates the CDF curve of the entropy. CDF curve towards the bottom-right is better. Results show that our model achieves higher entropy and is more uncertain about the out-of-distribution images.

The results on the above uncertainty metrics demonstrate the superiority of our Bayesian model about the uncertainty estimates even if we change to another discriminative model.

Classification Accuracy. Our Bayesian model achieves better classification accuracy than the GNN model with 50.80 ± 0.34 *v.s.* 50.33 ± 0.36. This results show that our model can achieve not only better uncertainty estimation but also better classification accuracy than the GNN model.

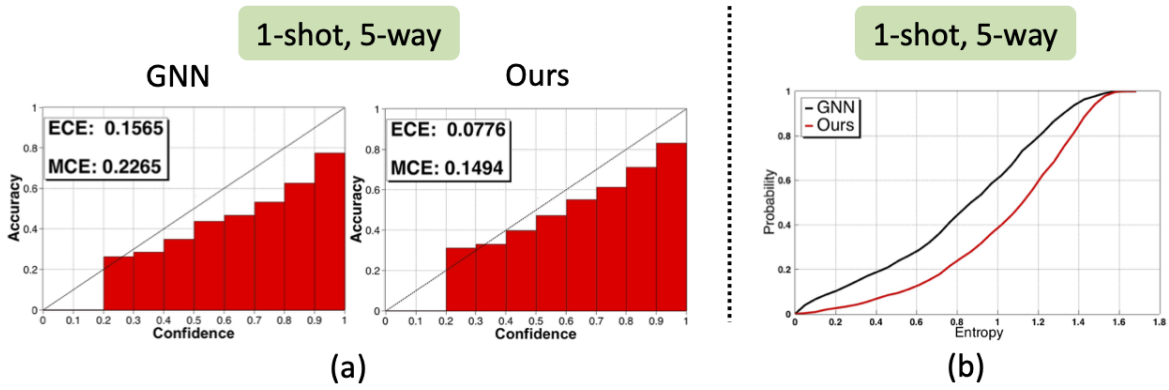

Fig. 5.6 (a) Comparison on the model calibration. (b) Comparison on out-of-distribution images.

5.5 Conclusion

This dissertation proposes a metric-based Bayesian framework to address the few-shot classification problem. Compared with MAML-based Bayesian methods, our method directly models the discriminative model to predict the query set given support set, without task-specific optimization. We employ a natural-gradient VI method to optimize the objective function. We use the ProtoNet as the discriminative model. However, our proposed Bayesian framework is generic and easy to extend to other metric-based few-shot learning models. Extensive experiments are conducted and our model achieves comparable or better performance than previous state-of-the-arts, including both the classification accuracy and the uncertainty estimation. Moreover, our model is more robust against attacks and prevents overfitting.

Chapter 6

Conclusion

Recently, deep learning models are adequately developed for vision applications. In this dissertation, I mainly focus on two limitations of the deep models. First, most early works employ the global representation extracted by deep models to tackle the vision problems and ignore the discriminative partial representations. Second, deep learning models are eager for massive labeled data, which is inefficient or expensive to acquire. Thus, the lack of labeled data will introduce uncertainty inherently. A robust model needs to not only predict accurate results but also reason about the uncertainty. This dissertation presents the part-aware and robust deep learning-based models for three vision applications, the occluded person re-id, the interactive video object segmentation, and the few-shot image classification.

For the occluded person re-id, I propose to partition the feature map into partial features, and use pose keypoints to indicate the visible parts and the occluded parts. The visible keypoints are utilized to generate the spatial attention maps and re-calibrate the channel responses for the feature map. During inference, the part-aware features in the common visible region between the query and gallery images are used to compute the distances.

For the interactive video object segmentation (VOS), I propose a part-aware and unified framework for both interaction and transference. The information is passed to the present processing frame from the interactive frame and the previous frame. I first extract the pixel embeddings and compute the distance of pixel embeddings in a local part or the global map. The global and local maps are utilized for generating the segmentation masks. Besides, I propose memory modules to aggregate the information from previous interaction rounds.

For the few-shot image classification, a metric-based Bayesian framework is presented in the dissertation. I adopt a discriminative model that takes as input the support set

to predict the label of the query set, without optimization during inference. I employ a natural-gradient VI method to approximate the posterior distribution of the network parameters.

In the future, I will continue focusing on the part-aware models for vision applications, especially for video applications. Objects in adjacent frames show the similarity in parts. Thus, an excavation about the partial similarity is necessary to generate more robust and discriminative representations, which will benefit the down-stream vision tasks.

References

[1] An, L., Chen, X., Yang, S., and Li, X. (2016). Person re-identification by multi-hypergraph fusion. *IEEE transactions on neural networks and learning systems*, 28(11):2763–2774.

[2] Bao, L., Wu, B., and Liu, W. (2018). Cnn in mrf: Video object segmentation via inference in a cnn-based higher-order spatio-temporal mrf. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5977–5986.

[3] Benard, A. and Gygli, M. (2017). Interactive video object segmentation in the wild. *arXiv preprint arXiv:1801.00269*.

[4] Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015a). Weight uncertainty in neural network. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1613–1622, Lille, France. PMLR.

[5] Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015b). Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*.

[6] Caelles, S., Maninis, K.-K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., and Van Gool, L. (2017). One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 221–230.

[7] Caelles, S., Montes, A., Maninis, K.-K., Chen, Y., Van Gool, L., Perazzi, F., and Pont-Tuset, J. (2018). The 2018 davis challenge on video object segmentation. *arXiv preprint arXiv:1803.00557*.

[8] Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2018). OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*.

[9] Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2019). Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186.

[10] Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*.

[11] Cardelli, L., Kwiatkowska, M., Laurenti, L., and Patane, A. (2019). Robustness guarantees for bayesian inference with gaussian processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7759–7768.

[12] Chang, X., Hospedales, T. M., and Xiang, T. (2018). Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2109–2118.

[13] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018a). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818.

[14] Chen, W., Chen, X., Zhang, J., and Huang, K. (2017). Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 403–412.

[15] Chen, Y., Pont-Tuset, J., Montes, A., and Van Gool, L. (2018b). Blazingly fast video object segmentation with pixel-wise metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1189–1198.

[16] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[17] Fang, H.-S., Xie, S., Tai, Y.-W., and Lu, C. (2017). RMPE: Regional multi-person pose estimation. In *ICCV*.

[18] Fathi, A., Wojna, Z., Rathod, V., Wang, P., Song, H. O., Guadarrama, S., and Murphy, K. P. (2017). Semantic instance segmentation via deep metric learning. *arXiv preprint arXiv:1703.10277*.

[19] Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR.

[20] Finn, C., Xu, K., and Levine, S. (2018). Probabilistic model-agnostic meta-learning. *arXiv preprint arXiv:1806.02817*.

[21] Fu, Y., Wei, Y., Zhou, Y., Shi, H., Huang, G., Wang, X., Yao, Z., and Huang, T. (2019). Horizontal pyramid matching for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8295–8302.

[22] Garcia, V. and Bruna, J. (2017). Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043*.

[23] Ge, Y., Li, Z., Zhao, H., Yin, G., Yi, S., Wang, X., et al. (2018). Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *Advances in neural information processing systems*, pages 1222–1233.

[24] Goodfellow, I., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

[25] Gordon, J., Bronskill, J., Bauer, M., Nowozin, S., and Turner, R. (2019). Meta-learning probabilistic inference for prediction. In *International Conference on Learning Representations*.

[26] Gordon, J., Bronskill, J., Bauer, M., Nowozin, S., and Turner, R. E. (2018). Meta-learning probabilistic inference for prediction. *arXiv preprint arXiv:1805.09921*.

[27] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330. JMLR. org.

[28] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[29] He, L., Liang, J., Li, H., and Sun, Z. (2018a). Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7073–7082.

[30] He, L., Sun, Z., Zhu, Y., and Wang, Y. (2018b). Recognizing partial biometric patterns. *arXiv preprint arXiv:1810.07399*.

[31] He, L., Wang, Y., Liu, W., Zhao, H., Sun, Z., and Feng, J. (2019). Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8450–8459.

[32] Heo, Y., Koh, Y. J., and Kim, C.-S. (2019a). Interactive video object segmentation using sparse-to-dense networks. In *CVPR Workshops*, volume 2, page 6.

[33] Heo, Y., Koh, Y. J., and Kim, C.-S. (2019b). Interactive video object segmentation using sparse-to-dense networks. *CVPR Workshops*.

[34] Hermans, A., Beyer, L., and Leibe, B. (2017). In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

[35] Hinton, G. (1979). Some demonstrations of the effects of structural descriptions in mental imagery. *Cognitive Science*, 3(3):231–250.

[36] Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347.

[37] Hu, J., Shen, L., and Sun, G. (2018a). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

[38] Hu, Y.-T., Huang, J.-B., and Schwing, A. G. (2018b). Videomatch: Matching based video object segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 54–70.

[39] Huang, C., Li, R., Li, X., and Fu, C.-W. (2020). Non-local part-aware point cloud denoising. *arXiv preprint arXiv:2003.06631*.

[40] Huang, H., Li, D., Zhang, Z., Chen, X., and Huang, K. (2018). Adversarially occluded samples for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

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

[42] Kalayeh, M. M., Basaran, E., Gökmen, M., Kamasak, M. E., and Shah, M. (2018). Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1062–1071.

[43] Khan, M., Nielsen, D., Tangkaratt, V., Lin, W., Gal, Y., and Srivastava, A. (2018). Fast and scalable Bayesian deep learning by weight-perturbation in Adam. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2611–2620, Stockholmsmässan, Stockholm Sweden. PMLR.

[44] Khan, M. E. and Lin, W. (2017). Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models. *arXiv preprint arXiv:1703.04265*.

[45] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[46] Kingma, D. P., Salimans, T., and Welling, M. (2015). Variational dropout and the local reparameterization trick. In *Advances in neural information processing systems*, pages 2575–2583.

[47] Krähenbühl, P. and Koltun, V. (2013). Parameter learning and convergent inference for dense random fields. In *International Conference on Machine Learning*, pages 513–521.

[48] Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

[49] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

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

[51] Li, W., Zhu, X., and Gong, S. (2017). Person re-identification by deep joint learning of multi-loss classification. *arXiv preprint arXiv:1705.04724*.

[52] Li, W., Zhu, X., and Gong, S. (2018). Harmonious attention network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[53] Liang, C., Yang, Z., Miao, J., Wei, Y., and Yang, Y. (2020). Memory aggregated cfbi+ for interactive video object segmentation. In *CVPR Workshops*, volume 1.

[54] Liao, S., Hu, Y., Zhu, X., and Li, S. Z. (2015). Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2197–2206.

[55] Liao, S., Jain, A. K., and Li, S. Z. (2012). Partial face recognition: Alignment-free approach. *IEEE Transactions on pattern analysis and machine intelligence*, 35(5):1193–1205.

[56] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

[57] Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Hu, Z., Yan, C., and Yang, Y. (2019). Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 95:151–161.

[58] Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., Yi, S., Yan, J., and Wang, X. (2017). Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pages 350–359.

[59] Liu, Y., Guo, J., and Lee, J. (2011). Halftone image classification using lms algorithm and naive bayes. *IEEE Transactions on Image Processing*, 20(10):2837–2847.

[60] Louizos, C. and Welling, M. (2017). Multiplicative normalizing flows for variational bayesian neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2218–2227. JMLR. org.

[61] Luiten, J., Voigtlaender, P., and Leibe, B. (2018). Premvos: Proposal-generation, refinement and merging for video object segmentation. In *Asian Conference on Computer Vision*, pages 565–580. Springer.

[62] Matsukawa, T., Okabe, T., Suzuki, E., and Sato, Y. (2016). Hierarchical gaussian descriptor for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1363–1372.

[63] Miao, J., Wei, Y., and Yang, Y. (2020). Memory aggregation networks for efficient interactive video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10366–10375.

[64] Miao, J., Wu, Y., Liu, P., Ding, Y., and Yang, Y. (2019). Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 542–551.

[65] Miao, J., Wu, Y., and Yang, Y. (2021). Identifying visible parts via pose estimation for occluded person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*.

[66] Naeini, M. P., Cooper, G., and Hauskrecht, M. (2015). Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[67] Najafi, M., Kulharia, V., Ajanthan, T., and Torr, P. (2018). Similarity learning for dense label transfer. *CVPR Workshops*.

[68] Neal, R. M. (2012). *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

[69] Oh, S. W., Lee, J.-Y., Sunkavalli, K., and Kim, S. J. (2018). Fast video object segmentation by reference-guided mask propagation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7376–7385.

[70] Oh, S. W., Lee, J.-Y., Xu, N., and Kim, S. J. (2019a). Fast user-guided video object segmentation by interaction-and-propagation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5247–5256.

[71] Oh, S. W., Lee, J.-Y., Xu, N., and Kim, S. J. (2019b). Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235.

[72] Oreshkin, B., López, P. R., and Lacoste, A. (2018). Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, pages 721–731.

[73] Osawa, K., Swaroop, S., Khan, M. E. E., Jain, A., Eschenhagen, R., Turner, R. E., and Yokota, R. (2019). Practical deep learning with bayesian principles. In *Advances in neural information processing systems*, pages 4287–4299.

[74] Pohlen, T., Hermans, A., Mathias, M., and Leibe, B. (2017). Full-resolution residual networks for semantic segmentation in street scenes. In *CVPR*, pages 4151–4160.

[75] Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., and Van Gool, L. (2017). The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*.

[76] Quan, R., Dong, X., Wu, Y., Zhu, L., and Yang, Y. (2019). Auto-reid: Searching for a part-aware convnet for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3750–3759.

[77] Ravi, S. and Beatson, A. (2019a). Amortized bayesian meta-learning. In *ICLR (Poster)*.

[78] Ravi, S. and Beatson, A. (2019b). Amortized bayesian meta-learning. In *International Conference on Learning Representations*.

[79] Ravi, S. and Larochelle, H. (2017). Optimization as a model for fewshot learning. In *International Conference on Learning Representations*.

[80] Ristani, E., Solera, F., Zou, R., Cucchiara, R., and Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*.

[81] Satorras, V. G. and Estrach, J. B. (2018). Few-shot learning with graph neural networks. In *International Conference on Learning Representations*.

[82] Silver, T., Allen, K. R., Lew, A. K., Kaelbling, L. P., and Tenenbaum, J. (2020). Few-shot bayesian imitation learning with logical program policies. In *AAAI*, pages 10251–10258.

[83] Snell, J., Swersky, K., and Zemel, R. S. (2017). Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*.

[84] Su, C., Li, J., Zhang, S., Xing, J., Gao, W., and Tian, Q. (2017). Pose-driven deep convolutional model for person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 3960–3969.

[85] Suh, Y., Wang, J., Tang, S., Mei, T., and Lee, K. M. (2018). Part-aligned bilinear representations for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–419.

[86] Sun, Y., Xu, Q., Li, Y., Zhang, C., Li, Y., Wang, S., and Sun, J. (2019). Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 393–402.

[87] Sun, Y., Zheng, L., Deng, W., and Wang, S. (2017). Svdnet for pedestrian retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3800–3808.

[88] Sun, Y., Zheng, L., Yang, Y., Tian, Q., and Wang, S. (2018). Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496.

[89] Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., and Hospedales, T. M. (2018). Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208.

[90] Ventura, C., Bellver, M., Girbau, A., Salvador, A., Marques, F., and Giro-i Nieto, X. (2019). Rvos: End-to-end recurrent network for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5277–5286.

[91] Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., and Wierstra, D. (2016). Matching networks for one shot learning. *arXiv preprint arXiv:1606.04080*.

[92] Voigtlaender, P., Chai, Y., Schroff, F., Adam, H., Leibe, B., and Chen, L.-C. (2019). Feelvos: Fast end-to-end embedding learning for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9481–9490.

[93] Voigtlaender, P. and Leibe, B. (2017). Online adaptation of convolutional neural networks for video object segmentation. *arXiv preprint arXiv:1706.09364*.

[94] Wang, G., Yuan, Y., Chen, X., Li, J., and Zhou, X. (2018). Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282.

[95] Wang, W., Song, H., Zhao, S., Shen, J., Zhao, S., Hoi, S. C., and Ling, H. (2019a). Learning unsupervised video object segmentation through visual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3064–3074.

[96] Wang, Z., Xu, J., Liu, L., Zhu, F., and Shao, L. (2019b). Ranet: Ranking attention network for fast video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3978–3987.

[97] Wu, Y., Lin, Y., Dong, X., Yan, Y., Ouyang, W., and Yang, Y. (2018). Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5177–5186.

[98] Xiu, Y., Li, J., Wang, H., Fang, Y., and Lu, C. (2018). Pose Flow: Efficient online pose tracking. In *BMVC*.

[99] Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., and Huang, T. (2018). Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*.

[100] Xu, Z., Zhang, W., Ye, X., Tan, X., Yang, W., Wen, S., Ding, E., Meng, A., and Huang, L. (2020). Zoomnet: Part-aware adaptive zooming neural network for 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12557–12564.

[101] Yang, L., Wang, Y., Xiong, X., Yang, J., and Katsaggelos, A. K. (2018a). Efficient video object segmentation via network modulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6499–6507.

[102] Yang, X., Zhou, P., and Wang, M. (2018b). Person reidentification via structural deep metric learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30(10):2987–2998.

[103] Yang, Z., Wei, Y., and Yang, Y. (2020). Collaborative video object segmentation by foreground-background integration. In *European Conference on Computer Vision*, pages 332–348. Springer.

[104] Yoon, J., Kim, T., Dia, O., Kim, S., Bengio, Y., and Ahn, S. (2018). Bayesian model-agnostic meta-learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 7343–7353.

[105] Yu, Q., Chang, X., Song, Y.-Z., Xiang, T., and Hospedales, T. M. (2017). The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching. *arXiv preprint arXiv:1711.08106*.

[106] Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

[107] Zhang, J., Zhao, C., Ni, B., Xu, M., and Yang, X. (2019). Variational few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1685–1694.

[108] Zhang, X., Chen, Y., Zhu, B., Wang, J., and Tang, M. (2020). Part-aware context network for human parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8971–8980.

[109] Zhang, X., Luo, H., Fan, X., Xiang, W., Sun, Y., Xiao, Q., Jiang, W., Zhang, C., and Sun, J. (2017). Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*.

[110] Zhao, L., Li, X., Zhuang, Y., and Wang, J. (2017). Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 3219–3228.

[111] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015a). Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124.

[112] Zheng, W.-S., Gong, S., and Xiang, T. (2011). Person re-identification by probabilistic relative distance comparison. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[113] Zheng, W.-S., Li, X., Xiang, T., Liao, S., Lai, J., and Gong, S. (2015b). Partial person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4678–4686.

[114] Zheng, Z., Zheng, L., and Yang, Y. (2017a). A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1):1–20.

[115] Zheng, Z., Zheng, L., and Yang, Y. (2017b). Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*.

[116] Zheng, Z., Zheng, L., and Yang, Y. (2018a). A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1):13.

[117] Zheng, Z., Zheng, L., and Yang, Y. (2018b). Pedestrian alignment network for large-scale person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(10):3037–3045.

[118] Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y. (2017). Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*.

[119] Zhu, K., Guo, H., Liu, Z., Tang, M., and Wang, J. (2020). Identity-guided human semantic parsing for person re-identification. *arXiv preprint arXiv:2007.13467*.

[120] Zhuo, J., Chen, Z., Lai, J., and Wang, G. (2018). Occluded person re-identification. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.