Scale Adaptive Network for Partial Person Re-identification: Counteracting Scale Variance

HongYu Chen[1,2]
ChenHY@mail.nwpu.edu.cn
BingLiang Jiao[†1,2]
bingliang.jiao@mail.nwpu.edu.cn
LiYing Gao[1,2]
gaoliying@mail.nwpu.edu.cn
Peng Wang[†1,2]
peng.wang@nwpu.edu.cn

[1] School of Computer Science and
Ningbo Institute, Northwestern
Polytechnical University.
Xi'an, China

[2] National Engineering Laboratory for
Integrated Aero-Space-Ground-Ocean.
Xi'an, China

Abstract

Partial Person Re-identification (Partial ReID) is a challenging task which aims to match partially visible images with holistic images of the same pedestrian. One of the significant challenges of this task is scale misalignment between holistic and partial person images, which makes it difficult for models to adapt to the scale gaps of different images. Previous methods used pooling or convolutional layers with various sizes to extract features of different scales. However, it is essential to note that each person's image has specifically suitable feature extraction scales, and some scale features may be unnecessary or even detrimental. Based on this finding, an adaptive feature extraction paradigm could be more suitable for Partial ReID. To this end, we propose a novel Scale Adaptive Network (SANet) to dynamically extract scale-adaptive features to counteract scale variance. Specifically, we introduce an Adaptive Feature Enhancement module (AFE) to adaptively extract multi-scale features and address scale misalignment. Furthermore, since a partial image only contains a portion of body parts in holistic images, the body parts exclusive to holistic images could introduce noise for image matching. Thus, we utilize a segmentation head to indicate the available human parts in each image and use the common visible body parts for feature comparisons between images. Extensive experiments demonstrate the effectiveness of our SANet network, which achieves comparable performance on partial and holistic person ReID datasets. Our code is available on https://github.com/chenjiangniao/SANet

1 Introduction

Person re-identification (ReID) aims to match the same individual captured by non-overlapping cameras, widely used in video surveillance and criminal investigation. In recent years, with the development of deep learning and the publication of large-scale holistic person datasets,

[†] Corresponding authors.

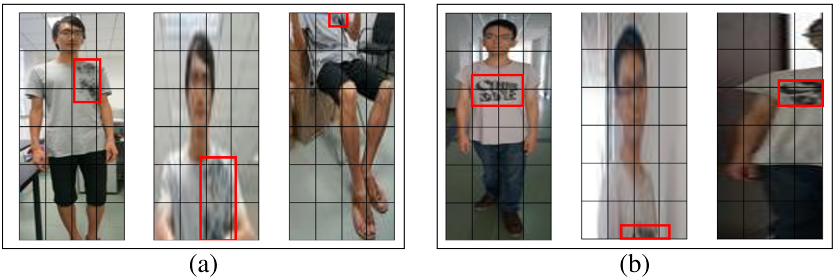


Figure 1: The scale variation between holistic and partial images. As denoted in the red boxes, the same local patterns in different images demonstrate significant visual variance, which requires ReID models to adapt to inputs with different scales during feature extraction.

holistic person ReID has demonstrated remarkable success. However, partial images are inevitably captured in real-world scenarios due to occlusion, camera range, and viewpoint. In these situations, directly applying holistic person ReID methods could lead to an inaccurate matching. Thus, partial person ReID was proposed in [39], which aims to match an identical person's partial and holistic images.

In partial person ReID, one of the main challenges comes from the scale variation between partial and holistic person images. As illustrated in Figure 1, identical local patterns, indicated by the red boxes, exhibit significant visual differences in various images. When a predefined model is employed to extract human features from these images, the image deformation caused by scale variation can pose substantial challenges in feature extraction, subsequently impacting feature comparison.

To address this problem, some methods propose locating the partial image within a holistic reference image using a learned position [8, 18]. However, this approach requires each compared image to serve as the reference image for a partial image to be located within. The pairwise comparison can be inefficient during the referencing stage, especially in a large image gallery. Alternatively, some methods [4, 5, 6, 17] design a spatial pyramid architecture to extract and fuse features of different scales, called multi-scale features, to combat scale misalignment. Although these methods can mitigate the scale misalignment issue by fusing multi-scale features, the predefined static architectures still lack adaptability to diverse scaled inputs encountered in real-world environments.

In this work, we introduce a novel network called *Scale Adaptive Network* (SANet) for Partial ReID. The network includes an adaptive feature extraction (AFE) module, which employs dynamic routing [14] to learn scale-adaptive features for partial person images with varying scales. Specifically, the AFE module takes the features from intermediate layers of ResNet-50 [3] as input and utilizes a series of scale-path selection cells to generate scale-adaptive features. Under the weak supervision of identity information, the AFE module learns to select the most suitable scale paths, resulting in more optimal feature representations.

Additionally, given that partial images contain only a portion of body parts compared to holistic images, the body parts exclusive to holistic images can introduce noise during image matching. To tackle this issue, we first employ human parsing masks to indicate the presence of each body part and extract the corresponding local features. Then, we utilize the features of commonly visible body parts to calculate the similarity between compared images.

The main contributions of our work can be summarised as follows:

- We propose an *Scale Adaptive Network* (SANet) for Partial ReID, which is designed to extract scale-adaptive features for partial images with arbitrary scales and extract shared local features for image matching.

- We propose an adaptive feature extraction (AFE) module with a series of scale-path selection and feature refinement cells. This module dynamically generates scale transformation paths and refines features layer by layer, ultimately enhancing their semantic richness.

- Our extensive experimental results demonstrate the effectiveness of the proposed AFE module. Moreover, the SANet network achieves comparable performance on partial, and holistic person ReID datasets.

2 Related Work

2.1 Holistic Person Re-Identification

Holistic person re-identification (ReID) focuses on matching images of the same pedestrian captured by different cameras. Existing ReID methods can be briefly classified into three categories, including feature representation learning methods, deep metric learning methods, and ranking optimization methods. Feature representation learning methods [25, 32, 37] are devoted to extracting robust and discriminative features from pedestrian images. Deep metric learning methods [19, 27, 29] try to design novel and effective loss functions to regulate the distance between positive and negative image pairs. The ranking optimization methods [1, 21, 34] introduce similarity ranking in the testing phase and re-optimizing the ranking list based on the similarities between images. In real-world scenarios, obstacles or limited camera range may result in images showing only a portion of a pedestrian, negatively impacting the performance of ReID models.

2.2 Partial Person Re-identification

Partial person re-identification (Partial ReID), proposed by [38], aims to match partial images with holistic images of the same pedestrians. The scale misalignment between partial and holistic images is one significant challenge for this task. To solve this problem, existing methods could be briefly divided into two categories [8], *i.e.*, multi-scale feature extraction methods [6, 8, 35] and feature reconstruction methods [4, 17, 26]. Multi-scale feature extraction methods aim to mitigate the effects of scale variance by extracting robust multi-scale features using pooling or convolutional layers of different sizes. For example, He *et al.* [6] leverage pyramid pooling to extract spatial pyramid features of different scales. Feature reconstruction methods aim to mitigate the effects of information asymmetry by extracting features common to both local and global images through image transformations or auxiliary models. Luo *et al.* [17] employ the predicting of affine parameters and samples the patches from the holistic images to match partial images. For Multi-scale feature extraction methods, existing methods of fixed structure have difficulty adapting to different scales. In this work, we propose a scale-adaptive network to tackle the scale misalignment problem by generating scale-adaptive features for each input image.

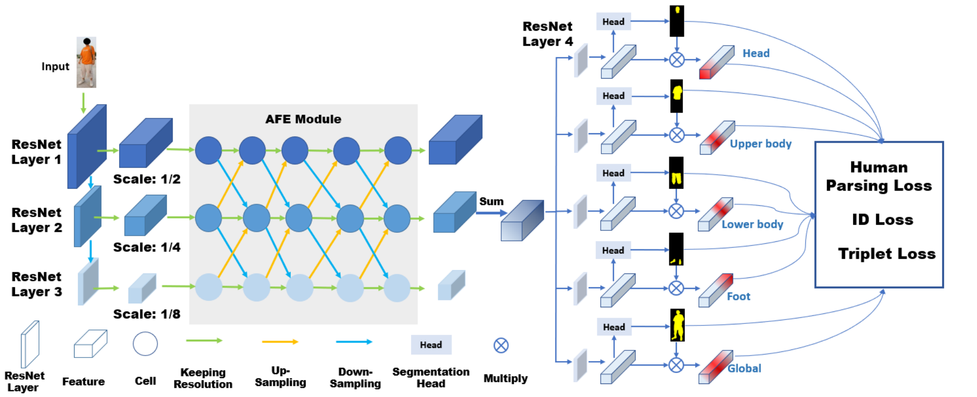


Figure 2: The overall architecture of our proposed SANet. It consists of an adaptive feature extraction (AFE) module and a multi-branch feature extraction (MFE) module. First, the AFE module adaptively generates dynamic multi-scale features against scale variations. Subsequently, we extract the visible information for matching.

2.3 Dynamic Routing

Dynamic routing is a dynamic network that adapts to samples of different scales by changing model structure. It aims to solve the problems like scale variance or limited computational resources [2, 13, 14]. Shen *et al.* [22] introduce dynamic routing in the field of medical image processing, which enhances feature maps by dense connections. Zhou *et al.* [41] introduce dynamic routing to visual question answering by dynamically selecting the attention span of the transformer. Considering that the ReID task is a real-time interactive task demanding high efficiency, our approach is more specific to Partial ReID. we thus choose to utilize the dynamic routing algorithm for extracting scale-adaptive features in this study.

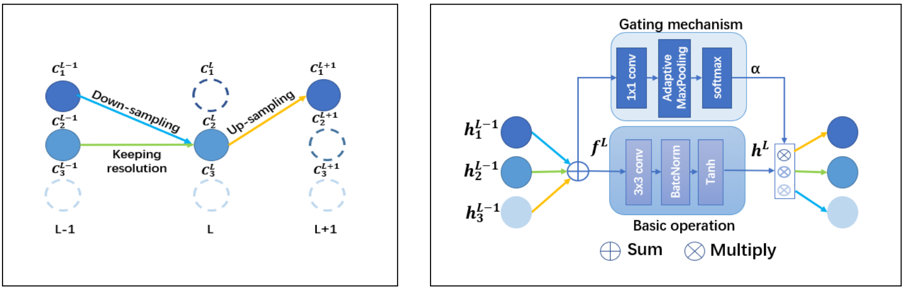
3 Methods

As illustrated in Figure 2, the SANet network consists of an adaptive feature enhancement (AEF) module and a multi-branch Feature Extension (MFE) module. The AFE module adaptively generates scale transformation paths for each image to obtain dynamic multi-scale features. The MFE module extracts local and global pedestrian features with the aid of human parsing masks and only calculates the similarity of common visible body parts between holistic and partial images.

3.1 Adaptive Feature Extraction Module

In Partial ReID, scale misalignment between holistic and partial images, as one of the significant challenges, brings considerable difficulty for a ReID model in extracting human features for image matching. To alleviate this problem, some previous methods use a spatial pyramid architecture to extract multi-scale features. However, this pre-defined static network still has difficulty adapting diverse scaled inputs.

In this work, we propose an adaptive feature enhancement (AEF) module, which utilizes dynamic routing [14] to generate scale-adaptive features dynamically. In terms of structure,



(a) (b)

Figure 3: (a) The interconnections between the various layers in AFE. (b) The construction of each cell. It aggregates the features from the h_i^{L-1} and generates scale transformation score α through a gating mechanism, and the basic operation further refines the features f^L to hidden features h^L.

we insert the AFE module between the first three layers and the 4-th layer of the backbone model ResNet-50 [3]. It comprises five layers containing three cells that produce different scale selections. The cells are interconnected between the layers to achieve scale transformations. After 5 layers of scale transformation selection, the generated multi-scale features are summarized together as the final scale-adaptive features for the image.

As illustrated in Figure 3 (a), c_2^L in L layer incorporates scale features from cells c_1^{L-1} and c_2^{L-1} in $L-1$ layer, refining the features at layer L. Subsequently, an up-sampling operation is applied to activate c_1^{L+1} in $L+1$ layer. As shown in Figure 3 (b), each cell in the AFE module is composed of three parts: scale transformation operation, gating mechanism, and base operation. The scale transformation operation includes up-sampling, down-sampling, and keeping resolution. In each cell, the exact process can be divided into four steps.

First, we summarize the features from the cells of the previous layer to get the aggregated feature $f^L = \sum_i^N h_i^{L-1}$, where h_i^{L-1} is the feature of the previous layer, and N represents the number of prior activated cells. Secondly, we design the gating mechanism to generate the probability of selecting different scale transformation paths, which comprises a 1×1 convolutional layer, a max-pooling layer, and a gumble softmax function [11]. It can be represented as follows:

$$\alpha^L = \mathrm{GS}(MaxPool((G_{1\times1}(w_i^L, f^L)))), \tag{1}$$

where α^L is the scale transformation score of i-th cell in the L-th layer, $G_{1\times1}$ is the 1×1 convolution operation, w_i^L represents the parameters of convolution layer, $MaxPool$ represents the maxpooling layer, GS represents the gumble softmax operation. Thirdly, the base operation consists of 3×3 convolution layer for feature refinement to improve the semantic richness. It can be represented as follows:

$$f_{refine}^L = G_{3\times3}(w_i^L, f^L), \tag{2}$$

where $G_{3\times3}$ is the 3×3 convolution operation, f_{refine}^L represents the features refined by base operation. Finally, we select the scale transformation paths using α, and perform the corresponding scale transformation, this process can be written as:

$$h_i^L = \alpha \times T_{scale}(f_{refine}^L), \tag{3}$$

where h_i^L denotes the final output of the cell, T_{scale} represent the corresponding scale transformation operation. In each layer, cells activated by the previous layer select the scale transformation paths to activate the cells of the next layer. We obtain scale transformation paths by passing and activating layer-by-layer and generating dynamic multi-scale features.

3.2 Multi-branch Feature Extension Module

Partial images, containing only sections of pedestrian figures, present an inherent information asymmetry compared to holistic images, significantly complicating image matching. Therefore, when computing similarities between image pairs, focusing on the features of commonly visible human body parts becomes crucial. To discern the presence of each body part and accurately extract the corresponding part-level features, we utilize human parsing masks as auxiliary information in this work.

In order to reduce the computational cost, we do not use an additional pre-trained human parsing model to generate human parsing masks, as in [10, 23]. Instead, we use 5 parameter-unshared branches to generate the masks for the human body parts (including the head, upper body, lower body, and feet) and the whole human content, as shown in Figure 2. Here, the global branch is utilized as a residual path. In each branch, a block initialized with the 4-th layer of ResNet-50 is first utilized. Then, a specific segmentation head, which consists of two convolutional layers, is utilized to produce human parsing masks. To optimize the segmentation heads, a human parsing loss is utilized, where the annotations are obtained from [12]. After that, the features for each branch can be calculated by:

$$F^i = \mathrm{MaxPool}(m^i \times f^i_{Res4}), \tag{4}$$

where $f^i_{Res4} \in \mathbb{R}^{h\times w\times C}$ represents the feature of the 4-th layer of ResNet-50 extracted from the i-th branch, $m^i \in \mathbb{R}^{h\times w\times 1}$ and $F^i \in \mathbb{R}^C$ denote the mask and the generated feature of the i-th branch, and MaxPool means a max-pooling layer. Besides, following [10, 23], the visible coefficient v^i for each body part can be calculated by,

$$v^i = max(0, sum(m^i > 0.5)) \geq 1, \tag{5}$$

where *sum* means the summation of pixel by pixel. After generating the visual feature and visible coefficient v^i for each body part, we can compute the similarities between images with the features of commonly visible body parts.

3.3 Training and Test

In the training stage, three loss functions, including human parsing loss, identity classification (ID) loss, and triplet loss [9], are utilized to optimize the SANet network. Specifically, we follow [10] to use focal loss [16] $\mathcal{L}_{Focal}$ as the human parsing loss to obtain the masks for each body part. Besides, following [10, 23, 25], we employ ID loss $\mathcal{L}_{ID}$ and triplet loss $\mathcal{L}_{Tri}$ on the features of each branch to teach the model to re-identity. The overall loss functions can be represented as,

$$\mathcal{L} = \sum_{i=1}^{5} \left(\mathcal{L}_{ID}(f_p^i) + \mathcal{L}_{Tri}(f_p^i) + \mathcal{L}_{Focal}(m^i, \mathrm{DS}(M^i))\right), \tag{6}$$

where $M^i \in \mathbb{R}^{H\times W\times 1}$ denotes the mask label for i-th branch, and DS means down-sampling to produce mask labels with the same size as m^i. We set the weight of triplet loss to 0.1, the other losses' to 1.

In the test phase, we choose the scale transformation path with the max probability in Equation 1 instead of using Gumbel-Softmax [11] to generate the scale-adaptive features for each input. We take the extracted F^i and use the v^i obtained by Equation 5 to determine whether the current local or global features are visible, and only calculate the Euclidean distance of visible parts. It can be represented as,

$$D = \frac{\sum_{i=1}^{5} \mathbb{1}^i \cdot Dist(F_q^i, F_g^i)}{\sum_{i=1}^{5} \mathbb{1}^i}, \quad \mathbb{1}^i = \begin{cases} 1, v_q^i \geq 1 \& \; v_g^i \geq 1, & v_q^i \geq 1 \; and \; v_g^i \geq 1 \\ 0, \; else. \end{cases} \tag{7}$$

where F_q^i,F_g^i represent the features of the i-th local branch belonging to the query image and gallery image, v_q^i and v_g^i mean the corresponding visible coefficient, and $\mathbb{1}^i$ is an indicator. We only calculate the distance between images when the body parts are both available.

4 Experiments

4.1 Datasets and Experiments Protocol

Partial-REID [38] dataset comprises 600 photographs of 60 distinct pedestrians, each with five holistic and five partial body images. **Partial-iLIDS** [4] is a dataset designed for partial person re-identification and contains 476 photos of 119 individuals. The partial-iLIDS dataset consists of two distinct forms [31]: Partial-iLIDS-O and Partial-iLIDS-P. Partial-iLIDS-O includes the barriers, whereas Partial-iLIDS-P crops the barriers to concentrate exclusively on pedestrians. **Market1501** [36] contains 1501 IDs which are captured by six different cameras, the training set contains 750 identities, and the test set contains 751 identities. **DukeMTMC-reID** [40] consists 16522 training images of 702 identities, 2228 query images of the other 702 identities and 17661 gallery images.

Experiment protocol. Following the standard evaluation setting, we adopt mean Average Precision (mAP) and Rank-k accuracy as the experimental protocol.

4.2 Implementation Details

As the Partial-REID and Partial-iLIDS are very small, we follow the previous works [4, 6, 8, 26] to train the network on the Market1501 [36] and test on the Partial-REID and Partial-iLIDS. We trained on two NVIDIA 3090 TI with 150 epochs, We scaled all photos to 384 × 128, then enhanced the data with flipping, color improvement, and random cropping. Before the training stage, following [24], we train our network with SGD optimizer, setting the learning rate to 0.01 and batch size to 64. We set the weight of triplet loss to 0.1, the other losses' to 1

4.3 Comparison with State-of-the-art Methods

Comparisons on Partial Person ReID Datasets. To exhibit the effectiveness of our proposed SANet, we compare the ReID performance of our SANet with the holistic, occluded and partial person ReID methods on the Partial-REID, Partial-iLIDS-O and Partial-iLIDS-P datasets. The experimental results are given in Table 1. From this table, we can find that our SANet achieves state-of-the-art Rank-1 and mAP performance on these three datasets. Compared to the best competitor FRT [30], our SANet outperforms by 6.2% in Rank-1 on

Category	Method	Partial-REID			Partial-iLIDS-O			Partial-iLIDS-P		
		Rank-1	Rank-3	mAP	Rank-1	Rank-3	mAP	Rank-1	Rank-3	mAP
Holistic	PCB (ECCV18) [25]	66.3	-	63.8	-	-	-	-	-	-
	TransReID (ICCV21) [7]	71.3	-	68.6	-	-	-	-	-	-
Occluded	FPR (ICCV19) [6]	81.0	-	76.6	-	-	-	68.1	-	61.8
	HPNet (ICME2018) [10]	85.7	-	81.8	72.0	-	58.9	68.9	80.7	72.2
	LKWS (ICCV21) [31]	85.7	**93.7**	-	80.7	88.2	-	-	-	-
	ASAN (TCVST21) [28]	86.8	93.5	78.8	81.7	88.3	85.9	71.4	81.9	72.5
	PAT (CVPR21) [15]	88.0	92.3	-	-	-	-	-	-	-
	FRT (TIP22) [30]	88.2	93.2	-	73.0	87.0	-	-	-	-
Partial	DSR (CVPR18) [4]	53.7	72.3	-	-	-	-	55.5	68.0	-
	VPM (CVPR19) [26]	67.7	83.6	-	-	-	-	67.2	76.5	-
	STNReID (TMM20) [17]	66.7	80.3	-		-	-	54.6	71.3	-
	PMN (AI22) [20]	76.7	79.0	-	-	-	-	62.2	74.8	-
	FSA (ICME22) [33]	73.7	82.7	-	-	-	-	68.9	82.4	-
	PPCL (CVPR21) [8]	83.7	88.7	-	-	-	-	71.4	**85.7**	-
	SANet (Ours)	**88.7**	92.3	81.5	**84.9**	**89.1**	**86.5**	**74.8**	84.0	**77.3**

Table 1: Performance comparison with state-of-the-art methods on three Partial ReID datasets, *i.e*., Partial-REID, Partial-iLIDS-O and Partial-iLIDS-P.

Method	Market1501		DukeMTMC	
	Rank-1	mAP	Rank-1	mAP
DSR [4]	83.5	64.2	-	-
VPM [26]	93.0	80.8	83.6	**72.6**
PGFA [18]	91.2	76.8	82.6	65.5
STNReID [17]	**93.8**	84.9	-	-
SANet (Ours)	93.7	80.1	**85.5**	67.1

Table 2: Performance comparison with state-of-the-art methods on two holistic person ReID datasets, *i.e*., Market1501 and DukeMTMC datasets.

average. The reason could be that our employed scale-adaptive feature enhancement strategy could adaptively generate effective features for inputs with variance scales.

Comparisons on Holistic Person ReID Datasets. To evaluate the effectiveness of our SANet on holistic person ReID, we also conduct experiments on two holistic person ReID datasets, *i.e*., Market1501 and DukeMTMC. The experimental results are exhibited in Table 2. Although the scale variation problem in holistic images is not as salient as in partial images, our SANet still achieves compatible performance on both holistic datasets, which indicates that our scale adaptive strategy is also effective for holistic images.

4.4 Ablation Studies

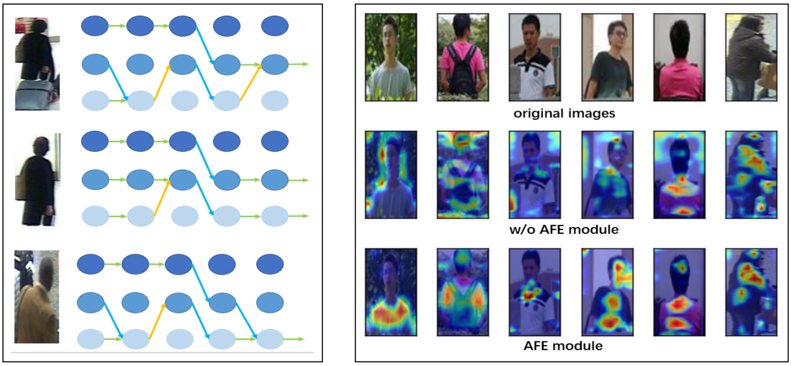
Effectiveness of the Components in SANet network. To demonstrate the effectiveness of the modules of SANet network, we conduct ablation experiments on the Partial-iLIDs-P dataset. The experimental results are exhibited in Table 3. For comparison, we first give the results of the baseline model, *i.e*., HPNet [10], which uses ResNet-50 as the backbone

	AFE	MFE	Rank-1	Rank-3	Rank-5	mAP
HPNet [10]	×	Local	68.9	80.7	82.4	72.2
+ Global Branch	×	✓	72.3	81.5	**85.7**	75.3
+ Fixed FE	Fixed	✓	73.9	79.0	83.2	76.3
SANet	✓	✓	**74.8**	**84.0**	84.9	**77.3**

Table 3: The effectiveness of the AFE module and MFE module.

Layer Number	Rank-1	mAP	Cell Operation	Rank-1	mAP
3	**74.8**	76.8	Conv1 × 1	73.1	75.9
5	**74.8**	**77.3**	Conv3 × 3	**74.8**	**77.3**
7	72.3	75.2	Conv5 × 5	73.1	75.6

Table 4: Ablation studies about the number of layers in the AFE module, and the base operations in a cell.



(a) (b)

Figure 4: (a) The visualization of different inference paths of instances in AFE. (b) The comparison of activation maps generated by models using or not using AFE module.

and four local branches. Then we additionally employ a global branch in the multi-branch feature extraction module, which improves the Rank-1 accuracy and mAP by 3.4% and 3.1%. Additionally, we use a fixed version of the feature enhancement (FE) module where the model is static and all scale transition paths are activated. It brings an improvement of 1.6% in Rank-1 and 1.0% in mAP, which means multi-scaled features are effective for this task. Furthermore, a scale-adaptive FE module, *i.e.*, AFE module, is utilized in our SANet, which further improves the Rank-1 accuracy and mAP by 0.7%/1.0%, respectively. The performance improvement shows the effectiveness of adaptive path transformation selection for scale variation.

Analysis of the deployment of our AFE module. Here, we analyze the impact of different layers on AFE modules. In the left part of Table 4, we set the number of layers as 3, 5, 7. The 5-layer of AFE performs best, with both too deep and too shallow layers detrimental to the dynamic scale transformation.

Analysis about the convolution sizes in AFE module. We compare different base operations of the cell, setting the size of the convolution kernel from 1×1 to 5×5. The experimental results are shown in the right part of Table 4. A larger convolution kernel tends to miss small information, while a smaller kernel is hard to capture more comprehensive information. According to the results, we set the convolution kernel to be 3×3.

4.5 Qualitative Analysis

In this section, we present the qualitative experimental results and demonstrate the superiority of our SANet. Figure 4 (a) exhibits that the AFE module generates adaptive inference paths for different inputs. For the holistic pedestrian images, the shallow features are extracted to prevent the loss of detailed information by continuous down-sampling. For the partial images, the deep features are extracted to alleviate large scaling. Besides, we demonstrate the activation maps of our AFE module in Figure 4 (b). It shows that the AFE module can adapt to different scales of pedestrians. Our AFE module is able to extract finer pedestrian areas and adaptively extract pedestrian features.

5 Conclusion

In this paper, we propose a novel *Scale Adaptive Network* (SANet) for Partial ReID. Our SANet addresses the scale misalignment problem by adaptively extracting dynamic multi-scaled features and refining them layer by layer, we also utilize a multi-branch feature extraction module to mitigate the information asymmetry between comparing partial and holistic images. Experiments demonstrate that our SANet achieves comparable performance on partial and holistic person ReID datasets.

Acknowledgement

This work is supported by National Key R&D Program of China (No.2020AAA0106900), the National Natural Science Foundation of China (No.U19B2037), Shaanxi Provincial Key R&D Program (No.2021KWZ–03), Natural Science Basic Research Program of Shaanxi (No.2021JCW–03).

References

[1] Song Bai, Peng Tang, Philip H.S. Torr, and Longin Jan Latecki. Re-ranking via metric fusion for object retrieval and person re-identification. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 740–749, 2019. doi: 10.1109/CVPR.2019.00083.

[2] An-Chieh Cheng, Chieh Hubert Lin, Da-Cheng Juan, Wei Wei, and Min Sun. Instanas: Instance-aware neural architecture search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3577–3584, 2020.

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

[4] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. 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, 2018.

[5] Lingxiao He, Zhenan Sun, Yuhao Zhu, and Yunbo Wang. Recognizing partial biometric patterns. *CoRR*, abs/1810.07399, 2018. URL http://arxiv.org/abs/1810.07399.

[6] Lingxiao He, Yinggang Wang, Wu Liu, He Zhao, Zhenan Sun, and Jiashi Feng. 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, 2019.

[7] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022, 2021.

[8] Tianyu He, Xu Shen, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. Partial person re-identification with part-part correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9105–9115, 2021.

[9] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[10] Houjing Huang, Xiaotang Chen, and Kaiqi Huang. Human parsing based alignment with multi-task learning for occluded person re-identification. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020.

[11] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[12] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. OpenPifPaf: Composite fields for semantic keypoint detection and spatio-temporal association. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):13498–13511, 2022. doi: 10.1109/TITS.2021.3124981.

[13] Yanwei Li, Lin Song, Yukang Chen, Zeming Li, Xiangyu Zhang, Xingang Wang, and Jian Sun. Learning dynamic routing for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8553–8562, 2020.

[14] Yanwei Li, Lin Song, Yukang Chen, Zeming Li, Xiangyu Zhang, Xingang Wang, and Jian Sun. Learning dynamic routing for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8553–8562, 2020.

[15] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2898–2907, 2021.

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

[17] Hao Luo, Wei Jiang, Xing Fan, and Chi Zhang. Stnreid: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification. *IEEE Transactions on Multimedia*, 22(11):2905–2913, 2020.

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

[19] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016.

[20] Qilu Qiu, Jieyu Zhao, and Ye Zheng. Partial person re-identification using a pose-guided alignment network with mask learning. *Applied Intelligence*, 52(10):10885–10900, 2022.

[21] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 420–429, 2018.

[22] Yan Shen and Mingchen Gao. Dynamic routing on deep neural network for thoracic disease classification and sensitive area localization. In *International Workshop on Machine Learning in Medical Imaging*, pages 389–397. Springer, 2018.

[23] Vladimir Somers, Christophe De Vleeschouwer, and Alexandre Alahi. Body part-based representation learning for occluded person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1613–1623, 2023.

[24] Vladimir Somers, Christophe De Vleeschouwer, and Alexandre Alahi. Body part-based representation learning for occluded person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1613–1623, 2023.

[25] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 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, 2018.

[26] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. 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, 2019.

[27] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6398–6407, 2020.

[28] Guanshuo Wang, Xiong Chen, Jialin Gao, Xi Zhou, and Shiming Ge. Self-guided body part alignment with relation transformers for occluded person re-identification. *IEEE Signal Processing Letters*, 28:1155–1159, 2021.

[29] Xinshao Wang, Yang Hua, Elyor Kodirov, Guosheng Hu, Romain Garnier, and Neil M Robertson. Ranked list loss for deep metric learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5207–5216, 2019.

[30] Boqiang Xu, Lingxiao He, Jian Liang, and Zhenan Sun. Learning feature recovery transformer for occluded person re-identification. *IEEE Transactions on Image Processing*, 31:4651–4662, 2022.

[31] Jinrui Yang, Jiawei Zhang, Fufu Yu, Xinyang Jiang, Mengdan Zhang, Xing Sun, Ying-Cong Chen, and Wei-Shi Zheng. Learning to know where to see: a visibility-aware approach for occluded person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11885–11894, 2021.

[32] Hantao Yao, Shiliang Zhang, Richang Hong, Yongdong Zhang, Changsheng Xu, and Qi Tian. Deep representation learning with part loss for person re-identification. *IEEE Transactions on Image Processing*, 28(6):2860–2871, 2019.

[33] Kai Zhang, Zheyang Li, Haoji Hu, Bin Li, Wenming Tan, Haixian Lu, Jun Xiao, Ye Ren, and Shiliang Pu. Dynamic feature pyramid networks for detection. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2022. doi: 10.1109/ICME52920.2022.9859874.

[34] Xuanmeng Zhang, Minyue Jiang, Zhedong Zheng, Xiao Tan, Errui Ding, and Yi Yang. Understanding image retrieval re-ranking: a graph neural network perspective. *arXiv preprint arXiv:2012.07620*, 2020.

[35] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. Pyramidal person re-identification via multi-loss dynamic training. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8506–8514, 2019. doi: 10.1109/CVPR.2019.00871.

[36] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015. doi: 10.1109/ICCV.2015.133.

[37] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1367–1376, 2017.

[38] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4678–4686, 2015. doi: 10.1109/ICCV.2015.531.

[39] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 4678–4686, 2015.

[40] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3774–3782, 2017. doi: 10.1109/ICCV.2017.405.

[41] Yiyi Zhou, Tianhe Ren, Chaoyang Zhu, Xiaoshuai Sun, Jianzhuang Liu, Xinghao Ding, Mingliang Xu, and Rongrong Ji. Trar: Routing the attention spans in transformer for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2074–2084, 2021.