Multi-scale Context Aggregation for Video-Based Person Re-Identification

Lei Wu[1], Canlong Zhang[1,2(✉)], Zhixin Li[1,2], and Liaojie Hu[3]

[1] Key Lab of Education Blockchain and Intelligent Technology Ministry of Education, Guangxi Normal University, Guilin, China
[2] Guangxi Key Lab of Multi-source Information Mining and Security, Guangxi Normal University, Guilin, China
zcltyp@163.com
[3] The Experimental High School Attached to Beijing Normal University, Beijing, China

Abstract. For video-based person re-identification (Re-ID), effectively aggregating video features is the key to dealing with various complicated situations. Different from previous methods that first extracted spatial features and later aggregated temporal features, we propose a Multi-scale Context Aggregation (MSCA) method in this paper to simultaneously learn spatial-temporal features from videos. Specifically, we design an Attention-aided Feature Pyramid Network (AFPN), which can recurrently aggregate detail and semantic information of multi-scale feature maps from the CNN backbone. To enable the aggregation to focus on more salient regions in the video, we embed a particular Spatial-Channel Attention module (SCA) into each layer of the pyramid. To further enhance the feature representations with temporal information while extracting the spatial features, we design a Temporal Enhancement module (TEM), which can plug into each layer of the backbone network in a plug-and-play manner. Comprehensive experiments on three standard video-based person Re-ID benchmarks demonstrate that our method is competitive with most state-of-the-art methods.

Keywords: Video-based Person Re-identification · Multi-scale Feature Aggregation · Feature Pyramid

1 Introduction

Person Re-Identification (Re-ID) aims to retrieve pedestrian targets with the same identity from multiple non-overlapping cameras, which has a high practical application value in society and industry. Compared with the conventional image-based person Re-ID, video-based person Re-ID can obtain more pedestrian information (e.g., action information and viewpoint information), thus alleviating the negative influence of the common occlusion situation in person Re-ID. Therefore, video-based person Re-ID has begun to receive academic attention and develop rapidly.

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024
B. Luo et al. (Eds.): ICONIP 2023, CCIS 1968, pp. 98–109, 2024.
https://doi.org/10.1007/978-981-99-8181-6_8

Currently, many person Re-ID methods [10,13,17,25] take ResNet [3] as their backbone network to extract features, which can effectively avoid the gradient disappearance and explosion of deep neural networks. However, limited by the size of the receptive field and the pooling operation, some image information is inevitably lost during feature learning. Besides, ResNet focuses more on the local region in the image and lacks modeling the correlation between human parts. These weaknesses limit the person Re-ID ability. To alleviate this problem, the attention mechanism [11,20] has begun to be widely used in video-based person Re-ID and improved the model performance, demonstrating its powerful representation ability by discovering salient regions in images.

The higher-level features of ResNet are abundant in semantics but lacking of details, while the lower-level features have more details but not enough semantics, so previous works [1,16,27] explore the effectiveness of hierarchical features in ResNet for video-based person Re-ID. In fact, features with different levels can complement each other through a specific aggregation. The deeper the layer is, the smaller the scale of its feature map is. Naturally, the feature maps of all layers are stacked together like a feature pyramid, which allows feature aggregation for video-based person Re-ID. However, effectively aggregating features of different layers through the pyramid structure is crucial for dealing with various complicated situations. PANet [14] proposed a bidirectional Feature Pyramid Network (FPN) consisting of a top-down as well as a bottom-up path to aggregate features at each layer of FPN. Similarly, Bi-FPN [9] proposed a nonlinear way to connect high-level and low-level. M2det [24] adopted a block of alternating joint U-shape module to fuse multilevel features. These methods have brought some improvements but require a large number of parameters and computations with complex structure.

In this paper, we innovatively embed the attention mechanism and Gated Recurrent Unit (GRU) into FPN to propose a learning method called Multi-scale Context Aggregation (MSCA), which can recurrently aggregate detail and semantic information of multi-scale feature maps from the backbone by Attention-aided FPN (AFPN), and enable the aggregation to focus on more salient regions in video. Different from previous methods that first extracted spatial features and later aggregated temporal features, our method can aggregate the spatio-temporal features simultaneously due to our proposed Temporal Enhancement Module (TEM). The TEM takes GRU as a primary component, it can be plugged into anywhere in a plug-and-play manner to learn complementary clues in temporal dimension and trained and converged easily due to its fewer parameters.

Overall, the main contributions of this paper are summarized as follows:

- We propose an Attention-aided Feature Pyramid Network (AFPN) to recurrently aggregate the hierarchical features from each layer of the backbone in spatial and channel dimensions; thus, the network can exploit salient clues and avoid some wrong clues during aggregation.

- We propose a Temporal Enhancement Module (TEM) that can be plugged into the backbone in a plug-and-play manner to merge the spatial and temporal information from pedestrian videos.
- Based on AFPN and TEM, we propose a Multi-scale Context Aggregation (MSCA) method for video-based person Re-ID. We conduct extensive experiments on three widely used benchmarks (i.e., MARS, iLIDS-VID and PRID2011), the results demonstrate that our MSCA method is competitive with other state-of-the-art methods, and the ablation studies confirm the effectiveness of AFPN and TEM.

2 Method

2.1 Overview

The overall architecture of our proposed method MSCA is illustrated in Fig. 1. We first adopt ResNet-50 [3] as our backbone and use multi-scale features from Res2, Res3, Res4, and Res5. Then, we plug TEM into the above four stages in a plug-and-play manner for learning temporal information from video features, which makes the features output from each layer contain both spatial and temporal information. We propagate the hierarchical features recurrently in AFPN for context aggregation, combining high-level semantic information with low-level detail information, and the SCA module would be plugged after aggregation to focus on more salient regions for improving the model performance. Finally, we use cross entropy loss and triplet loss as the objective function to optimize the model in the training stage, and features from two aggregate directions are concatenated for testing.

2.2 Attention-Aided Feature Pyramid Network

In video-based person Re-ID, multi-scale information aggregation has been used as one of the main methods to improve performance. The hierarchical features are characterized by insufficient semantic information of low-level features and insufficient detail information of high-level features due to the fact that the ResNet backbone increases the feature dimensions and decreases the feature resolutions across contiguous layers, thus Lin et al. [12] use this intrinsic property to reverse the information aggregation to form FPN based on the backbone by top-down and lateral connection. High-level features with rich semantic information are up-sampled by nearest neighbor interpolation and aggregated with the next layer of features output by lateral connection through element-wise addition, where the lateral connection adopts 1×1 convolution layer for reducing channel dimensions, and then recurrently aggregates swallow features, which take on more detail information and less semantic information. A 3×3 convolution layer is utilized on each aggregated features to generate the final feature map to reduce the aliasing effect of the upsampling operation. Besides, in order to make the aggregated features more discriminative, we introduce a spatial-channel

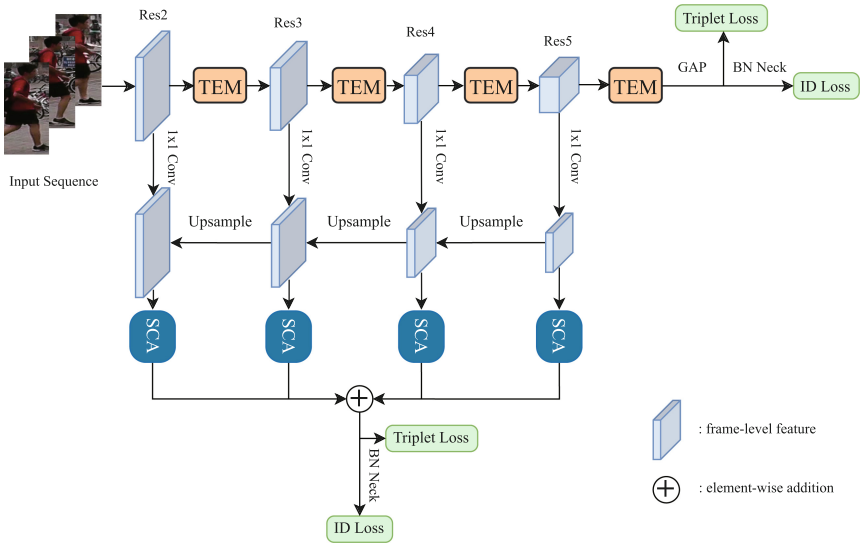

Fig. 1. Illustration of our proposed Multi-scale Context Aggregation. The input frames first fed to the ResNet-50 backbone. For the multi-scale feature maps from each residual block, we propose the AFPN to recurrently aggregate the features and focus on more salient regions by the SCA module. TEM is our proposed temporal enhancement module, which can be plugged into anywhere of the model for learning spatio-temporal information simultaneously. Finally, multiple losses are used to supervise the model in the training stage.

attention (SCA) module in FPN and call it attention-aided FPN (AFPN), the features of each layer after aggregation are fed to the SCA module, which consists of spatial and channel attention. As shown in Fig. 2, two kinds of attention cascade to calculate the spatial attention and channel attention [19], which can be described as follows:

$$F_s = A_s(F) \otimes F \tag{1}$$

$$F_{sc} = A_c(F_s) \otimes F \tag{2}$$

Where $\otimes$ denotes the element-wise multiplication, given the input feature tensor $F \in \mathbb{R}^{C\times H\times W}$, $A_s(\cdot)$ and $A_c(\cdot)$ denote the computation of the spatial and channel attention maps, $F_{sc} \in \mathbb{R}^{C\times H\times W}$ is the final spatial-aided and channel-aided features.

To obtain spatial attention, we adopt two pooling operations GAP and GMP to generate two features: $F^s_{avg} \in \mathbb{R}^{1\times H\times W}$ and $F^s_{max} \in \mathbb{R}^{1\times H\times W}$, then the two are concatenated to obtain a two-layer feature descriptor. The feature map is processed using a 7×7 convolution layer and a sigmoid layer to generate a spatial attention map $A_s(F) \in \mathbb{R}^{1\times H\times W}$, it is calculated as follows:

$$A_s(F) = \sigma(conv_{7\times 7}([F^s_{max}, F^s_{avg}])) \tag{3}$$

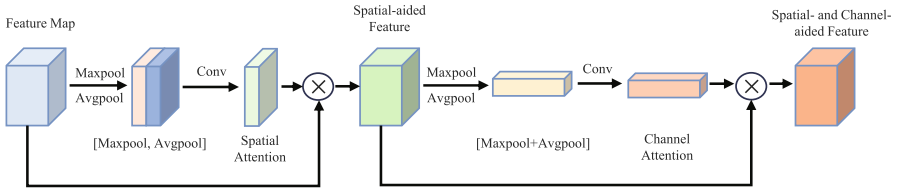

Fig. 2. Diagram of the SCA module

The channel attention helps the model to focus on more salient features by assigning greater weights to channels that show a higher response. We also first adopt GAP and GMP to generate two features: $F^c_{avg} \in \mathbb{R}^{C\times1\times1}$ and $F^c_{max} \in \mathbb{R}^{C\times1\times1}$, and the two are added up to fed into a convolution layer, which contains one ReLU activation layer and two 3×3 convolution layers. After that, the channel attention map $A_c(F) \in \mathbb{R}^{C\times1\times1}$ can be formulated as follows:

$$A_c(F) = \sigma(conv_{3\times3}([F^s_{max} + F^s_{avg}])) \tag{4}$$

To eliminate the superimposed effect of distracted information in weighted process, the final video feature is then obtained by the SCA module instead of primitive 3×3 convolution layer.

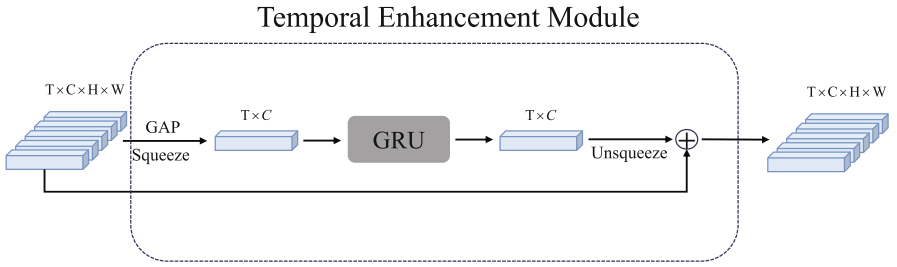

Fig. 3. Detailed structure of our TEM

2.3 Temporal Enhancement Module

Video sequences involve rich temporal information, we design a Temporal Enhancement Module (TEM) based on GRU, and it can be plugged anywhere in the model to capture the temporal clues in the video feature maps in a plug-and-play manner. As shown in Fig. 3, we use GAP to the input feature map $F \in \mathbb{R}^{T\times C\times H\times W}$ and squeeze all the spatial information to obtain temporal vector $F' \in \mathbb{R}^{T\times C}$. In the next, the temporal vector would be processed by GRU and recover the outputted feature map to the same size as the original input, then adopt a skip connection with input to obtain the final temporal enhanced

feature $F'' \in \mathbb{R}^{T\times C\times H\times W}$, which incorporates the temporal information in the video and the video feature representation is more discriminative. The whole process of TEM is formulated as follows:

$$F' = Squeeze(GAP(F)) \tag{5}$$

$$F'' = Unsqueeze(GRU(F')) + F \tag{6}$$

2.4 Loss Function

We employ two kinds of losses to jointly supervise the training the model: cross entropy loss and hard triplet loss [4]. We calculated two losses L_{xent} and L_{htri} as follows:

$$L_{xent} = \sum_{i=1}^{N} -q_i \log(p_i) \tag{7}$$

$$L_{htri} = [d_{pos} - d_{neg} + m]_+ \tag{8}$$

Where p_i is the predicted logit of identity i and q_i is the ground-truth label in identification loss, d_{pos} and d_{neg} are respectively defined as the distance of positive sample pairs and negative sample pairs, $[d]_+ = max(\cdot, 0)$ and m is the distance margin, which is set to 0.3 in the training procedure.

In this paper, in order to better supervise the training of model, we adopt output from Res5 in the backbone and output from AFPN in the two directions of feature aggregation. Therefore, the total loss is the summation of the four losses:

$$L_{total} = L_{xent}^{b2t} + L_{htri}^{b2t} + L_{xent}^{t2b} + L_{htri}^{t2b} \tag{9}$$

3 Experiments

3.1 Datasets

MARS [26] dataset is the biggest video-based person re-identification benchmark with 1261 identities and around 20000 video sequences generated by DPM detector and GMMCP tracker. The dataset is captured by six cameras, each identity is captured by at least two cameras and has 13.2 sequences on average. There are 3248 distracter sequences in the dataset, it increases the difficulty of Re-ID.

iLIDS-VID [10] dataset is captured by two cameras in an airport hall. It contains 600 videos from 300 identities. This benchmark is very challenging due to pervasive background clutter, mutual occlusions, and lighting variations.

PRID2011 [5] dataset captures 385 identities by camera A and 749 identities by camera B, but only the first 200 people appear in both cameras.

Table 1. Comparison with State-of-the-Art methods On MARS, iLIDS-VID and PRID2011 datasets.

model	Ref	MARS				iLIDS-VID			PRID2011		
		Rank-1	Rank-5	Rank-20	mAP	Rank-1	Rank-5	Rank-20	Rank-1	Rank-5	Rank-20
ADFD [25]	CVPR2019	87.0	95.4	98.7	78.2	86.3	97.4	99.7	93.9	99.5	100
GLTR [10]	ICCV2019	87.0	95.8	98.2	78.5	86.0	98.0	-	95.5	100	-
MG-RAFA [23]	CVPR2020	88.8	97.0	98.5	85.9	88.6	98.0	99.7	95.9	99.7	100
TCLNet [7]	ECCV2020	88.8	-	-	83.0	86.6	-	-	-	-	-
MGH [21]	CVPR2020	90.0	96.7	98.5	85.8	85.6	97.1	99.5	94.8	99.3	100
SSN3D [8]	AAAI2021	90.1	96.6	98.0	86.2	88.9	97.3	98.8	-	-	-
BiCnet-TKS [6]	CVPR2021	90.2	-	-	-	86.0	-	-	-	-	-
GRL [15]	CVPR2021	91.0	96.7	98.4	84.8	90.4	98.3	99.8	96.2	99.7	100
CTL [13]	CVPR2021	91.4	96.8	98.5	86.7	89.7	97.0	100	-	-	-
GPNet [17]	NN2022	90.2	96.8	98.8	85.1	88.8	98.5	100	96.1	99.8	100
PiT [22]	TII2022	90.2	97.2	-	86.8	92.1	98.9	100	-	-	-
Ours		**91.8**	96.5	98.1	83.2	84.7	94.7	**100**	**96.6**	**100**	**100**

3.2 Evaluation Metrics

We employ the Cumulative Matching Characteristic curve (CMC) and the mean Average Precession (mAP) as evaluation critiria. CMC considers re-ID as a ranking problem and represents the accuracy of the person retrieval with each given query, mAP reflects the true ranking results while multiple ground-truth sequences exist. Conveniently, Rank-1, Rank-5 and Rank-20 are employed to represent the CMC curve.

3.3 Implementation Details

ResNet-50 pre-trained on ImageNet is employed as our backbone, and the input images are all resized to 256×128. We also utilize some commonly data augmentation strategies including random horizontal flipping, random erasing and random cropping. Specifically, in the training stage, we employ a restricted random sampling strategy to randomly sample $T = 8$ frames from each video as input. The ADAM optimizer with an initial learning rate of 5×10^{-5} and a weight decay of 5×10^{-4} for updating the parameters. We train the model for 200 epochs, and the learning rate is reduced by 0.1 per 50 epochs. All the experiments are conducted with Pytorch and a NVIDIA RTX 3090 GPU.

3.4 Comparison with State-of-the-Art Methods

In this section, we compare our proposed method with other state-of-the-art video-based person Re-ID methods on MARS, iLIDS-VID and PRID2011.

Results on MARS. From Table 1, compared with other state-of-the-art methods, the proposed MSCA achieves the best Rank-1 accuracy and competitive Rank-5 and Rank-20 accuracy. According to the results given in CTL baseline [13] and our baseline in Table 2, although both of them adopt ResNet-50 as the

backbone, the mAP score of our baseline is 78.3% and that of CTL baseline is 82.7%, thus there is still a large margin in the final mAP result, even though our method performs well. The new and best works for video-based person Re-ID, CTL and PiT [22], the former adopts ResNet-50 as the backbone, but the latter is based on Transformer [2], all of which have utilized some complex modules such as key-points estimator, topology graph learning and hard-to-train transformer. In comparison, our method reaches a best Rank-1 accuracy with effective context aggregation. This demonstrates that our MSCA can aggregate more discriminative information, and the brief feature learning structure also has good generalization performance.

Results on iLIDS-VID and PRID2011. We also conduct several experiments on the two small datasets to demonstrate the advantages and possible flaws of our proposed method over the existing methods as shown in Table 1. We can observe that the result for iLIDS-VID is worse than other state-of-the-art methods, which causes this result is that the video sequences have a large variation of light and the serious occlusions on iLIDS-VID dataset. TEM only considers the temporal correlation but ignores the low-quality video sequences, which will introduce some additional irrelevant information, so as to decreases the model performance. For PRID2011 on the same scale as iLIDS-VID, our method achieves the best Rank-1 accuracy of 96.6%, and outperforms all previous approaches, confirming the superiority of our proposed method.

3.5 Ablation Study

To demonstrate the effectiveness of our proposed methods, we perform ablation studies on MARS dataset and use strong CNN backbone with ResNet-50 as our baseline. The experimental results are reported in Table 2 and Table 3.

Table 2. Ablation analysis of two components on MARS dataset

Model	AFPN	TEM	Rank-1	Rank-5	mAP
Baseline	✗	✗	88.2	95.4	78.3
	✓	✗	89.1	96.1	79.4
	✗	✓	90.7	96.5	82.1
Ours	✓	✓	**91.8**	**96.5**	**83.2**

MSCA: As shown in Table 2, the baseline contains only the ResNet-50 backbone and is supervised by L_{xent}^{b2t} and L_{htri}^{b2t}, the Rank-1 and mAP accuracy of the baseline are 88.2% and 78.3%, respectively. We find that using AFPN to aggregate multi-scale diverse features recurrently, the corresponding performance is

89.1% in Rank-1 and 79.4% in mAP, which is attributed to the complementary information of high-level semantic features and low-level detail features. Moreover, using TEM alone, we can achieve 90.7% in Rank-1 and 82.1% in mAP, the result shows the temporal enhancement module can complement the individual spatial features to learn more temporal information in videos. Eventually, by adding AFPN and TEM to the baseline, the model can further learn more discriminative features with the effective multi-scale context aggregation method.

AFPN: To explore the effectiveness of the SCA module and its position in the FPN, we conduct some experiments under the setting of utilizing the TEM. Table 3 shows the comparison of different plugging positions and the performance gap between our proposed AFPN and vanilla FPN, "w/o SCA" denotes vanilla FPN without SCA module, "1×1" denotes plugging the SCA module into the lateral connection and "Up-sample" denotes plugging the SCA module into the Up-sampling process. With the SCA module after propagation, we can find that the Rank-1 accuracy and mAP score are improved by 0.5% and 0.8%, respectively. Meanwhile, depending on the plugging position of the SCA module, the effects are also different. In Table 3, we can observe that plugging the SCA module into the lateral connection and into the downward propagation, they both have different performance degradation compared to plugging the SCA module after propagation.

Table 3. Ablation analysis of different plugging positions on MARS dataset

Position	Rank-1	Rank-5	mAP
w/o SCA	91.3	**96.7**	82.4
1×1	91.1	96.1	82.0
Up-sample	91.0	96.6	82.4
Ours	**91.8**	96.5	**83.2**

3.6 Visualization

As shown in Fig. 4, we report examples of different identities with Grad-CAM [18], which is commonly used in computer vision for a visual explanation. To verify the effectiveness of our proposed MSCA, we compare the Grad-CAM visualization of the baseline with our method, and the three example images are selected at intervals of at least 10 frames in three independent sequences of the MARS dataset. The frames contain various conditions such as motion and partial occlusions. As shown in Fig. 4(a), the features extracted from the ordinary frames by our proposed method can capture more information, including torso, legs and accessory that is discriminative to the target person. In Fig. 4(b), our method can focus on the area with more motion compared to the baseline if the

subject has significant motion. As shown in Fig. 4(c), the features extracted by our proposed method can capture more body regions without additional bicycle information, which will enhance the representation of the target person. In general, our MSCA can effectively capture more spatio-temporal information and avoid some occlusions to improve the performance.

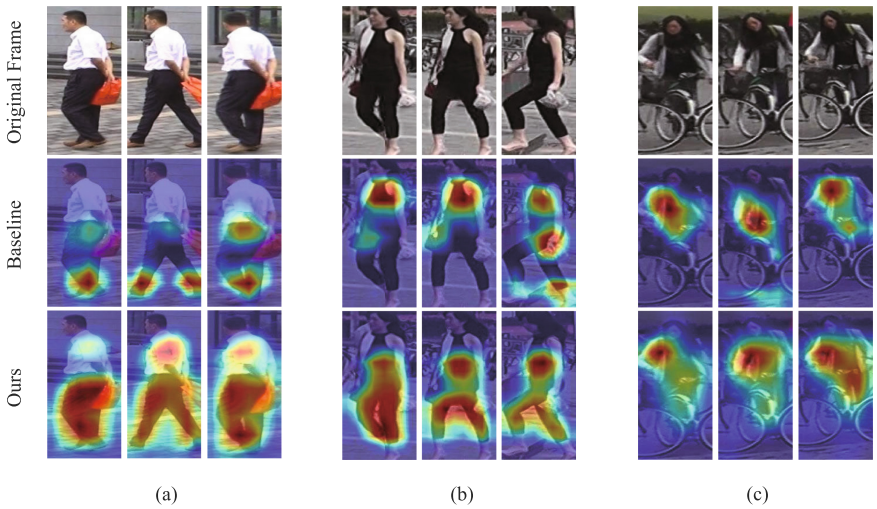

Fig. 4. Visualization of attention maps on different identities of the baseline and our proposed method. (a) Person takes up most of the image. (b) Person with significant motion. (c) Person is half occluded by bicycle

4 Conclusion

In this paper, we propose an innovative multi-scale context aggregation method for video-based person Re-ID. The proposed method can learn more video context information recurrently. AFPN can aggregate the semantic and detail information in multi-scale features, it integrates high-level semantic information into low-level detail information and uses the SCA module to aid the aggregated features to focus on salient regions. Furthermore, we propose a TEM to capture the temporal information among the video frames, and with its plug-and-play property, we can aggregate temporal features while extracting spatial features to enrich the final video feature representations, which is entirely different from previous works. The experimental results on three standard benchmarks demonstrate that our proposed method achieves competitive performance with most state-of-the-art methods.

Acknowledgements. This work is supported by National Natural Science Foundation of China (Nos. 62266009, 62276073, 61966004, 61962007), Guangxi Natural

Science Foundation (Nos. 2019GXNSFDA245018, 2018GXNSFDA281009, 2018GXNSFDA294001), Guangxi Collaborative Innovation Center of Multi-source Information Integration and Intelligent Processing, Innovation Project of Guangxi Graduate Education (YCSW2023187), and Guangxi "Bagui Scholar" Teams for Innovation and Research Project.

References

1. Chen, X., et al.: Salience-guided cascaded suppression network for person re-identification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3297–3307 (2020)
2. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv:2010.11929 (2020)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2015)
4. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv:1703.07737 (2017)
5. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Scandinavian Conference on Image Analysis (2011)
6. Hou, R., Chang, H., Ma, B., Huang, R., Shan, S.: BiCnet-TKS: learning efficient spatial-temporal representation for video person re-identification. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2014–2023 (2021)
7. Hou, R., Chang, H., Ma, B., Shan, S., Chen, X.: Temporal complementary learning for video person re-identification. In: European Conference on Computer Vision (2020)
8. Jiang, X., Qiao, Y., Yan, J., Li, Q., Zheng, W., Chen, D.: SSN3D: self-separated network to align parts for 3d convolution in video person re-identification. In: AAAI Conference on Artificial Intelligence (2021)
9. Kong, T., Sun, F., bing Huang, W., Liu, H.: Deep feature pyramid reconfiguration for object detection. arxiv:1808.07993 (2018)
10. Li, J., Wang, J., Tian, Q., Gao, W., Zhang, S.: Global-local temporal representations for video person re-identification. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3957–3966 (2019)
11. Li, S., Bak, S., Carr, P., Wang, X.: Diversity regularized spatiotemporal attention for video-based person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 369–378 (2018)
12. Lin, T.Y., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 936–944 (2016)
13. Liu, J., Zha, Z., Wu, W., Zheng, K., Sun, Q.: Spatial-temporal correlation and topology learning for person re-identification in videos. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4368–4377 (2021)
14. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8759–8768 (2018)

15. Liu, X., Zhang, P., Yu, C., Lu, H., Yang, X.: Watching you: global-guided reciprocal learning for video-based person re-identification. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13329–13338 (2021)
16. Liu, Z., Zhang, L., Yang, Y.: Hierarchical bi-directional feature perception network for person re-identification. In: Proceedings of the 28th ACM International Conference on Multimedia (2020)
17. Pan, H., Chen, Y., He, Z.: Multi-granularity graph pooling for video-based person re-identification. Neural Netw.?: Off. J. Int. Neural Netw. Soc. **160**, 22–33 (2022)
18. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. Int. J. Comput. Vision **128**, 336–359 (2016)
19. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: convolutional block attention module. In: European Conference on Computer Vision (2018)
20. Wu, L., Wang, Y., Gao, J., Li, X.: Where-and-when to look: deep Siamese attention networks for video-based person re-identification. IEEE Trans. Multimed. **21**, 1412–1424 (2018)
21. Yan, Y., et al.: Learning multi-granular hypergraphs for video-based person re-identification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2896–2905 (2020)
22. Zang, X., Li, G., Gao, W.: Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval. IEEE Trans. Industr. Inf. **18**(12), 8776–8785 (2022). https://doi.org/10.1109/TII.2022.3151766
23. Zhang, Z., Lan, C., Zeng, W., Chen, Z.: Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10404–10413 (2020). https://doi.org/10.1109/CVPR42600.2020.01042
24. Zhao, Q., et al.: M2det: a single-shot object detector based on multi-level feature pyramid network. In: AAAI Conference on Artificial Intelligence (2018)
25. Zhao, Y., Shen, X., Jin, Z., Lu, H., Hua, X.: Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4908–4917 (2019)
26. Zheng, L., et al.: Mars: a video benchmark for large-scale person re-identification. In: European Conference on Computer Vision (2016)
27. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3701–3711 (2019)