A Hybrid of Hard and Soft Attention for Person Re-Identification

Xuesong Li
The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences University of Chinese Academy of Sciences
Beijing, China
lixuesong2017@ia.ac.cn

Yating Liu
The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences University of Chinese Academy of Sciences
Beijing, China
liuyating2015@ia.ac.cn

Kunfeng Wang*
College of Information Science and Technology, Beijing University of Chemical Technology
Beijing, China
wangkf@mail.buct.edu.cn

Yong Yan
State Grid Zhejiang Electric Power Company, LTD
Hangzhou, China
yanyong@zj.sgcc.com.cn

Fei-Yue Wang
The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences
Beijing, China
feiyue.wang@ia.ac.cn

***Abstract*—Existing pedestrian re-identification methods based on deep learning have achieved good results under constrained conditions. However, there exist some challenges including large human pose variations, viewpoint changes, severe occlusions and imprecise detection of persons. So we present a Hard/Soft hybrid Attention Network (HSAN) that combines pose information and attention mechanism to deal with the challenges. Our model includes two main parts: Pose-guided Hard Attention (PHA) and Regional Soft Attention (RSA). PHA uses the keypoints generated by pose estimation to enhance the foreground information, and RSA is learned to eliminate the background clutter. We extract reliable features and locate discriminative regions by using these two modules to handle occlusions, pose changes and background noises. We conduct a lot of experiments on public datasets including DukeMTMC-ReID, Market-1501, and CUHK03, and the results show that our method achieves state-of-the-art performance.**

***Index Terms*—person re-identification, attention model, computer vision, deep learning.**

I. Introduction

Person re-identification (ReID) has become an essential task 5of computer vision. As described in Fig. 1, ReID is defined as using computer vision algorithms to find the same target under multiple cameras. Given query image and gallery set, ReID model ranks the gallery images based on the similarity score between gallery set and query set. ReID has many applications such as intelligent transportation, video surveillance, and human-computer interaction. Hence, it is imperative to implement thorough study on ReID.

ReID has attracted more and more attention. In general, there are two main approaches for ReID. One approach treats the ReID problem as a classification problem that is solved by predicting the ID of image, and is called representation learning [1]. Another approach aims to learn the similarity between two images through the network which is input image pairs, and is called discriminative distance metric learning [2]. In the early ReID research, lots of works focused on global feature for person retrieval. Later, it was found that global feature meets a bottleneck, because global feature ignores the fine-grained image characteristics. Recent articles have focused more on learning local feature and got better results [1], [3]. However, there are still many difficulties for the ReID task.

Fig. 1 shows the main problems of ReID. There are many

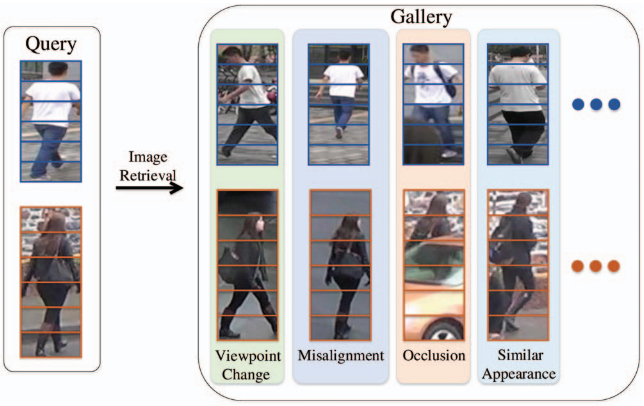

Fig. 1. The illustration and main challenges of person re-identification. The images in the first row are from Market-1501 dataset, and the images in the second row are from DukeMTMC-ReID dataset.

This work was supported by the Science and Technology Project of State Grid Corporation of China (Grant 52110418002Y).
Corresponding author: Kunfeng Wang.

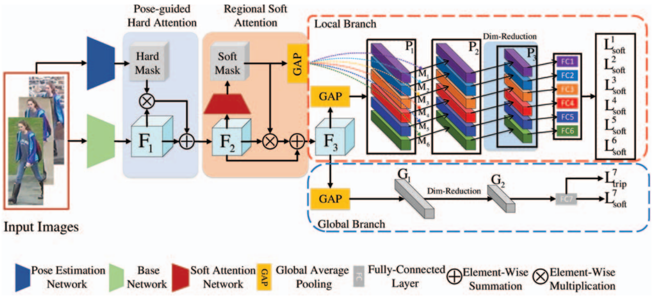

Fig. 2. The structure of HSAN. HSAN has two main parts: PHA module and RSA module. The PHA module aims to enhance the foreground features by producing Hard Mask. The RSA module aims to weaken the background noises by producing Soft Mask. F_1, F_2 and F_3 are the feature maps of each stage. F_3 is operated by Global Average Pooling (GAP) to obtain component features P_1. And the soft mask is also operated by GAP to get the regional attention masks. Each regional attention mask multiplies each stripe of component features P_1 to get feature maps P_2. F_3 is also operated by Global Average Pooling (GAP) to obtain global features G_1. Both P_2 and G_1 are operated respectively by dimension reduction to acquire P_3 and G_2. Finally all stripes of P_3 and G_2 are input respectively to a classifier.

human pose variations in the ReID datasets since person is non-rigid object. And because the ReID task is completed under multiple cameras, there will be different perspectives for the same target. The imprecise detection of persons will lead to pedestrian misalignment, which affects the performance of ReID. Some works try to solve these challenges by using semantic information such as human pose estimation [4], [5]. Although these methods lead to performance improvement to some degree, there are obvious issues, e.g., because pose informations are not fully utilized, there are still redundant background information and interferences in feature maps. There are also a few works that use attention mechanism to reduce the negative impacts from these challenges [6], [7]. Nevertheless, these works are unable to locate the discriminative regions and utilize the pixel-level information. To solve these problems, we should make full use of pose information and attention mechanism.

In this work, we make some improvements on feature extraction for ReID. We propose a Hard/Soft hybrid Attention Network (HSAN) that combines pose information and attention mechanism to deal with the above difficulties. HSAN consists of two main parts: Pose-guided Hard Attention (PHA) and Regional Soft Attention (RSA). PHA uses the keypoints generated by pose estimation to enhance the foreground information, and is utilized to calibrate poor detections. RSA is applied to weaken the background noise, and it is added to not only global branch but also part branch, taking into account the global-level and part-level attention at the same time. In order to verify the effectiveness of our method, we implement extensive experiments on three public datasets.

The remainders of this paper are organized as follows. In Section II, we propose a review on related works of ReID. In Section III, we describe the details of our unified framework and the proposed attention modules. In Section IV, we evaluate our model on multiple datasets and analyze our method by ablation study. Finally, the conclusion is draw in Section V.

II. Related Works

With the development of deep learning, some deep learning methods for ReID have achieved good results with high identification rates. These methods mainly include representation learning and discriminative metric learning. The representation learning method focuses on getting characteristics of pedestrians under different cameras, while the discriminative distance metric learning method focuses on learning the distance metric that maximizes matching accuracy.

Recent articles have focused not only on learning global feature that may omit pixel-level details, but also on learning local feature [1], [3]. Though part-based ReID models have achieved competitive performance, there are still many problemes to be solved. Some works introduce human pose estimation information to improve performance for ReID. For example, Xu et al. [4] present a Pose-guided Part Attention (PPA) module to mask out needless background noise. Wei et al. [5] present a Global-Local-Alignment Descriptor (GLAD) to generate discriminative character representation for solving misalignment and pose change problems. However, these methods only use semantic information to roughly separate foreground and background. They do not take into account the part-level attention and fine-grained information. Some works do not leverage cues from semantic parts. For example, Zheng et al. [6] present a Pedestrian Alignment Network (PAN) that makes use of attention mechanism to obtain reliable features. Zhang et al. [7] achieve the corresponding alignment of two pedestrians by computing the shortest distance between two local-feature sets. Nevertheless, these methods usually bring in redundant background noise to feature map, and result in inaccurate matches.

Inspired by the above methods, we propose an HSAN model to solve these difficulties. We present PHA module that uses pose feature to enhance foreground feature. We also present RSA module that uses soft attention mechanism to weaken the background information which is useless. We combine hard and soft attention so as to improve the performance of ReID.

III. Proposed Method

In this section, we mainly introduce the Hard/Soft hybrid Attention Network (HSAN) and its two components: PHA module and RSA module.

A. Hard/Soft Hybrid Attention Network (HSAN)

As described in Fig. 1, given the query image and gallery set, our task is to train a ReID model that is used to rank images of gallery set in accordance with similarity scores between gallery set and the query image.

In this work, we present HSAN to fulfil the ReID task. See Fig. 2, our backbone network mainly includes base network that is used to generate feature maps and pose estimation network that is used to obtain the keypoints of input images. Because ResNet-50 has relatively terse architecture and is widely used for ReID task, it is used as our base network.

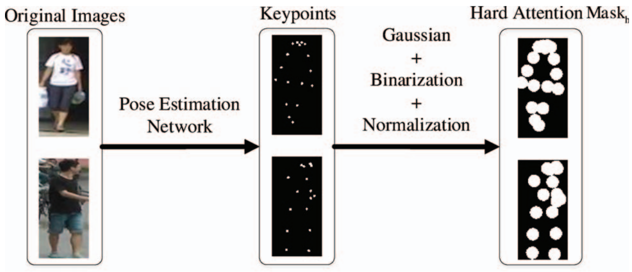

Fig. 3. The visualization of keypoints and hard attention. The keypoints are generated by pose estimation network. And hard attention mask is gained through the post-processing of keypoints.

And we make some minor modifications to the base network as in PCB [1], deleting the original GAP layer and the parts after GAP. In addition, we take no downsampling in the last convolutional layer. The base network is followed by PHA and RSA modules. The PHA module aims to enhance the reliability of features by using pose information. And the RSA module aims to locate the discriminative region by using attention mechanism. We describe the PHA module in Section III-B and RSA module in Section III-C in detail.

Feature map F_3 is operated by Global Average Pooling (GAP) to obtain global features G_1 and component features P_1. In particular, HSAN divides F_3 evenly into 6 horizontal stripes by using GAP. Afterwards, both P_2 in local branch and G_1 in global branch that are 2048-d are operated respectively by dimension reduction to acquire P_3 and G_2. Each stripe of P_3 or G_2 is set to 256-d and is input respectively to a classifier for predicting the target ID. And we implement all classifiers by FC layer and loss function.

B. Pose-guided Hard Attention Module

We use a pose estimation network to extract the pose information of pedestrians. And the Stacked Hourglass Network [8] is used as pose estimation network, which is trained on COCO dataset [9] to obtain 17 keypoints for ReID task. There exists pedestrian misalignment due to imprecise detection, which affects effectiveness of the part-based model. To weaken negative effect of misalignment, we employ keypoints to preprocess the input images.

In the PHA module, we also use keypoints to enhance the reliability of foreground features. Due to the distribution deviation between Re-ID and pose estimation datasets, and poor generalization performance of the pose estimation model across datasets, most methods [4] [5] that use pose information to divide pedestrian parts is not effective. Therefore, we consider to convert keypoint information into attention mechanism. When the pose estimation model fails, the performance of the model will not become worse. The specific implementation is shown as below. The coordinate (x_i, y_i) of keypoint K_i is mapped into the feature map F_1 that is generated by base network, which can be denoted by

$$(x_i', y_i') = (\lambda_1 x_i, \lambda_2 y_i), \tag{1}$$

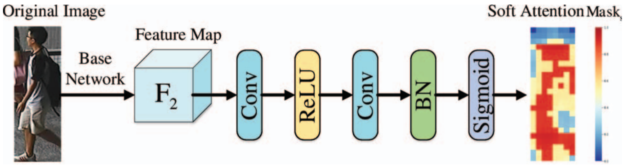

Fig. 4. Structure of the Soft Attention Network. It is part of Regional Soft Attention (RSA) module and is used to generate Soft Mask.

where λ_1 and λ_2 are the proportionality coefficient between the sizes of original images and feature maps. And (x_i', y_i') is the mapped location of keypoint K_i. Then the heat maps H are generated according to the mapped keypoints K', H_i is computed by the following formula

$$H_i = Binarization(Gaussian(K_i', \sigma), thres), \tag{2}$$

where $Gaussian$ is used to generate a Gaussian mask according to the central position K_i' and standard deviation σ of the Gaussian, and σ is set to 16. And $Binarization$ is applied to binarize the Gaussian mask with the threshold value $thres$, the $thres$ is set to 0.8. Because there are 17 keypoints, we can obtain 17 heat maps. The final Hard Mask $Mask_h$ is given by

$$Mask_h = Norm(\sum_{i=1}^{17} H_i), \tag{3}$$

where $Norm$ denotes the normalization operation for the summation of all heat maps H. We apply $Mask_h$ to mask feature map F_1, and then hard attention feature map F_2 is represented as

$$F_2 = F_1 \bigotimes Mask_h \bigoplus F_1, \tag{4}$$

where $\bigotimes$ and $\bigoplus$ denote element-wise multiplication and summation, respectively. We enhance information about regions of interest through the hard attention.

C. Regional Soft Attention Module

In the RSA module, we utilize attention mechanism to extract reliable features and locate discriminative regions. As shown in Fig. 4, the Soft Mask $Mask_s$ is computed by the following formula

$$Mask_s = Sigmoid(BN(Conv(ReLU(Conv(F_2))))), \tag{5}$$

where the two $Conv$ operators denote convolutional function, $ReLU$ and $Sigmoid$ refer to activation function, and BN refers to Batch Normalization. As described in Fig. 2, we apply Soft Mask to mask the feature map F_2,

$$F_3 = F_2 \bigotimes Mask_s \bigoplus F_2. \tag{6}$$

Soft Mask $Mask_s$ assigns weights to the channels of feature map F_2 according to preference of channels.

As Fig. 2 shows, Soft Mask is not only added to the backbone network, but also the network of local branch. For each stripe P_1^i of local feature maps P_1, the RSA module

TABLE I

COMPARISON BETWEEN STATE-OF-THE-ART METHODS AND OUR REID MODEL ON MARKET-1501 AND DUKEMTMC-REID DATASETS.

Method	Market-1501				DukeMTMC-reID			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
HA-CNN [17]	75.7	91.2	-	-	63.8	80.5	-	-
DaRe [18]	76.0	89.0	-	-	64.5	80.2	-	-
DuATM [19]	76.6	91.4	97.1	**99.0**	64.6	81.8	90.2	95.4
PABR [20]	79.6	91.7	96.9	98.1	69.3	84.4	92.2	93.8
DNN_CRF [21]	81.6	93.5	97.7	-	69.5	84.9	92.3	-
PCB+RPP [1]	81.6	93.8	97.5	98.5	69.2	83.3	90.5	92.5
Mancs [22]	82.3	93.1	-	-	71.8	84.9	-	-
SPReID [23]	83.4	93.7	97.6	98.4	73.3	86.0	93.0	94.5
Ours	**86.3**	**95.0**	**98.0**	98.6	**77.9**	**88.1**	**93.9**	**95.9**
DaRe (RR) [18]	86.7	90.9	-	-	80.0	84.4	-	-
PABR (RR) [20]	89.9	93.4	96.4	97.4	83.9	88.3	93.1	95.0
SPReID (RR) [23]	91.0	94.6	96.8	97.7	85.0	89.0	93.3	94.8
Ours (RR)	**93.4**	**95.4**	**97.6**	**98.2**	**89.3**	**91.7**	**94.9**	**96.4**

produces a regional soft mask M, which can be represented as

$$M = GAP(Mask_s, p), \tag{7}$$

where GAP refers to global average pooling operation. GAP splits $Mask_s$ into p regional soft masks. And each regional soft mask M_i is applied to each stripe of feature maps P_1 for generating local feature maps P_2,

$$P_2^i = M_i \bigotimes P_1^i. \tag{8}$$

The regional soft mask M_i is used to weaken the redundant background noise of local branch, helping improve the performance of our ReID model.

D. Implementation Details

We train pose estimation model on COCO dataset by using the Stacked Hourglass Network. The pose estimation model generates keypoints of pedestrians for our ReID task. We also crop input images to remove excess background information with the help of keypoints. Then we resize the cropped images to 384×128. And we initialize our base network with weighting coefficients of our base network pre-trained on ImageNet dataset [10].

We train our ReID model by using softmax and triplet losses. We use softmax loss to predict ID of pedestrian, which turns the ReID task into a classification problem. And it is formulated as

$$L_{soft} = -log\frac{e^{\omega_y}f}{\sum_{k=1}^{C} e^{\omega_k}f}, \tag{9}$$

where f is the learned feature of pedestrian, and ω_k denotes weight coefficient of class k. C indicates to the number of classes. Triplet loss aims to learn the similarity between the input image pairs, which is used for metric learning. And it is represented as

$$L_{trip} = -(\|f_a - f_p\|_2 - \|f_a - f_n\|_2 + \alpha), \tag{10}$$

where f_a, f_p and f_n are the learned features that are from anchor sample, positive sample and negative sample respectively. α is a margin parameter, and we set it to 1.2. For the local branch of HSAN, softmax loss is used for training our model. And for global branch, the final loss is got by adding softmax loss and triplet loss.

We use single GPU, set the batch size to 32 and use Adam optimizer to train our ReID model. Epoch is set to 500. We set initial learning rate to 2e-4, and we decay it to 2e-5 and 2e-6 progressively.

IV. EXPERIMENTS

In this part, we show the results of our method and compare its performance with state-of-the-art methods. And ablation experiment is conducted to show the availability of proposed attention modules.

A. Datasets and Protocols

Our ReID model and proposed attention modules are evaluated on Market-1501 [11], DukeMTMC-ReID [12] and CUHK03 [13] datasets. See Fig. 1, these datasets present lots of challenges including viewpoint change, misalignment, occlusion and similar appearance. For the CUHK03 dataset, we adopt the new division method which is more challenging than the original method [14]. The new protocol splits CUHK03 dataset, which consist of 767 identities for training set and 700 identities for testing set respectively.

For these datasets, we use the universal evaluation metrics that are widely employed in many ReID methods. We report mean Average Precision (mAP) [11] and Cumulated Matching Characteristics (CMC) [15] for rank-k accuracy to compare different approaches.

B. Results and Comparisons

Table I shows experimental results of our ReID method on Market-1501 and DukeMTMC-reID sets. We compare our method with the state-of-the-art methods for ReID. We report mAP and Rank-k accuracy for each method. “RR” denotes carrying out Re-Ranking. In these metrics, mAP and Rank-1 are the main evaluation metrics. As Table I shows, our ReID model realizes mAP/Rank-1=86.3%/95.0% on Market-1501 dataset and mAP/Rank-1=77.9%/88.1% on DukeMTMC-reID dataset. It is intuitive to see our ReID model is better than

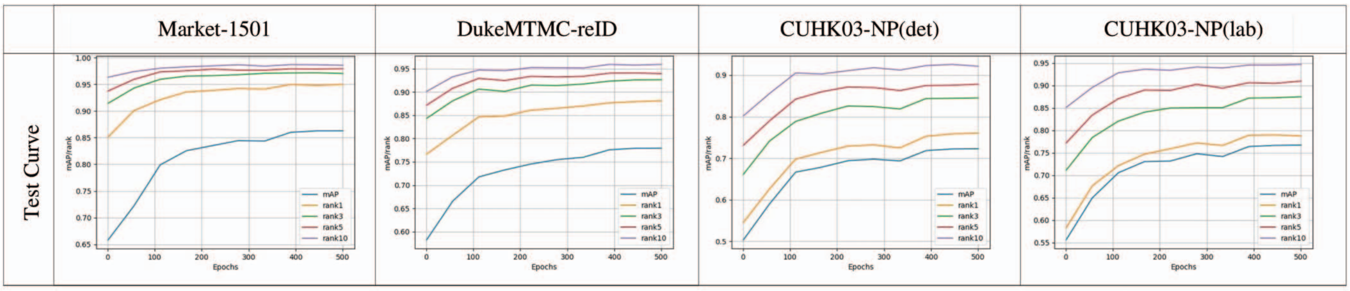

Fig. 5. The test curve of our ReID model that is trained on Market-1501, DukeMTMC-reID and CUHK03 datasets.

TABLE II
COMPARISON BETWEEN STATE-OF-THE-ART METHODS AND OUR REID MODEL ON CUHK03-NP DATASET.

Method	CUHK03-NP (det)		CUHK03-NP (lab)	
	mAP	Rank-1	mAP	Rank-1
HA-CNN [17]	38.6	41.7	41.0	44.4
MLFN [16]	47.8	52.8	49.2	54.7
TriNet [24]	50.7	55.5	53.8	58.1
PCB+RPP [1]	57.5	63.7	-	-
DaRe [18]	59.0	63.3	61.6	66.1
Mancs [22]	60.5	65.5	63.9	69.0
EANet [25]	66.8	72.5	-	-
Ours	**72.3**	**76.1**	**76.7**	**78.8**
Ours (RR)	**84.3**	**83.7**	**87.6**	**86.4**

TABLE III
EFFECTS OF THE PHA AND RSA MODULES. "BASELINE" DENOTES THE HSAN MODEL WITHOUT PHA AND RSA MODULES.

Method	Market-1501/DukeMTMC-reID			
	mAP	Rank-1	Rank-5	Rank-10
Baseline	84.9/74.8	93.9/86.0	97.9/92.6	**98.8**/94.7
Baseline+PHA	85.2/76.2	94.7/86.9	**98.0**/93.6	98.6/95.2
Baseline+RSA	**86.4**/75.3	94.7/86.6	**98.0**/92.7	98.7/94.8
HSAN	**86.4/77.9**	**95.0/88.1**	**98.0/93.9**	**98.8/95.9**

previous models on multiple evaluation criteria. We also compare the methods by implementing Re-Ranking [14] operation. The idea of Re-Ranking is based on the assumption that it is more possible to be a right match when the images of gallery set are similar to query image. The comparisons prove that our HSAN model is better than others on Market-1501 and DukeMTMC-reID sets.

Table II presents experimental results of ReID model on CUHK03-NP (det) and CUHK03-NP (lab) datasets. Our H-SAN model is also compared with state-of-the-art models. And we report mAP and Rank-1 for quantitative comparison. As Table II shows, our HSAN model achieves mAP/Rank-1=72.3%/76.1% on CUHK03-NP (det) dataset and mAP/Rank-1=76.7%/78.8% on CUHK03-NP (lab) dataset. Our HSAN model is shown to outperform other models on CUHK03 dataset.

Figure 5 show the test curve of our HSAN model that is trained on Market-1501, DukeMTMC-reID and CUHK03 sets. Each curve corresponds to the experimental results in the tables above. The abscissa of each graph is the training epochs. For the test curve, its ordinate is the value of each evaluative criteria including mAP, Rank-k.

C. Ablation Study

Our ablation experiments are implemented on Market-1501 and DukeMTMC-reID sets. For each ablation experiment, we strictly control all the hyper-parameters to be consistent. In the following, we verify the effectivenesses of our PHA module and RSA module through experimental results.

Table III shows the effects of the PHA and RSA modules on Market-1501 and DukeMTMC-reID sets. We mainly focus on mAP and Rank-1 metrics that are widely used for ReID task. The baseline model is our HSAN model without PHA and RSA modules. Compared to the results of baseline model, the PHA module achieves mAP/Rank-1=0.3%/0.8% rise on Market-1501 dataset and mAP/Rank-1=1.4%/0.9% rise on DukeMTMC-reID dataset. The PHA module produces more benefits on DukeMTMC-reID dataset than it does on the Market-1501 set. Because the DukeMTMC-reID set is more complex and noisy than Market-1501, it means that our PHA module is adaptive and robust to complex environments.

Compared to the results of baseline model, the RSA module achieves mAP/Rank-1=1.5%/0.8% rise on Market-1501 dataset and mAP/Rank-1=0.5%/0.6% rise on DukeMTMC-reID set. It indicates that our RSA module can improve effectiveness of ReID task. From Table III, we can further see that our HSAN model combining PHA and RSA modules performs better than using single one on most evaluation metrics. These experimental results prove HSAN method is effective.

V. CONCLUSION

In this paper, we present a Hard/Soft hybrid Attention Network (HSAN) that combines pose information and attention mechanism for ReID task. The HSAN model is a network structure that combines hard attention and soft attention. The hard attention produced by Pose-guided Hard Attention (PHA) module is used to enhance the foreground information. And soft attention produced by Regional Soft Attention (RSA) module is used to extract reliable feature and locate discriminative region. Extensive experiments on multiple datasets demonstrate performance of our method. In future work, we will apply our approach to more powerful CNN backbone structures such as ResNet-101 and DenseNet, and integrate the ReID model into other computer vision tasks, e.g., multi-target tracking.

REFERENCES

[1] Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond Part Models: Person Retrieval with Refined Part Pooling (And a Strong Convolutional Baseline). In Proceedings of the European Conference on Computer Vision (pp. 480-496) (2018)

[2] Xiao, Q., Luo, H., Zhang, C.: Margin Sample Mining Loss: A deep Learning Based Method for Person Re-Identification. arXiv preprint arXiv:1710.00478 (2017)

[3] Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning Discriminative Features with Multiple Granularities for Person Re-Identification. In 2018 ACM Multimedia Conference on Multimedia Conference (pp. 274-282). ACM (2018, October)

[4] Xu, J., Zhao, R., Zhu, F., Wang, H., Ouyang, W.: Attention-Aware Compositional Network for Person Re-Identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2119-2128. IEEE (2018)

[5] Wei, L., Zhang, S., Yao, H., Gao, W., Tian, Q.: GLAD: Global-Local-Alignment Descriptor for Pedestrian Retrieval. In: Proceedings of the 25th ACM international conference on Multimedia, pp. 420-428. ACM (2017, October)

[6] Zheng, Z., Zheng, L., Yang, Y.: Pedestrian Alignment Network for Large-Scale Person Re-Identification. IEEE Transactions on Circuits and Systems for Video Technology (2018)

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

[8] Newell, A., Yang, K., Deng, J.: Stacked Hourglass Networks for Human Pose Estimation. In European Conference on Computer Vision (pp. 483-499). Springer, Cham (2016, October)

[9] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L.: Microsoft COCO: Common Objects in Context. In European conference on computer vision (pp. 740-755). Springer, Cham (2014, September)

[10] Krizhevsky, A., Sutskever, I., Hinton, G. E.: Imagenet Classification with Deep Convolutional Neural Networks. In Advances in neural information processing systems (pp. 1097-1105) (2012)

[11] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable Person Re-Identification: A Benchmark. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1116-1124) (2015)

[12] Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance Measures and A Data Set for Multi-Target, Multi-Camera Tracking. In European Conference on Computer Vision (pp. 17-35). Springer, Cham (2016, October)

[13] Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep Filter Pairing Neural Network for Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 152-159) (2014)

[14] Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking Person Re-Identification with K-Reciprocal Encoding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1318-1327) (2017)

[15] Gray, D., Brennan, S., Tao, H.: Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. In Proceedings of the IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (Vol. 3, No. 5, pp. 1-7) Citeseer (2007, October)

[16] Chang, X., Hospedales, T. M., Xiang, T.: Multi-Level Factorisation Net for Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2109-2118) (2018)

[17] Li, W., Zhu, X., Gong, S.: Harmonious Attention Network for Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2285-2294) (2018)

[18] Wang, Y., Wang, L., You, Y., Zou, X., Chen, V., Li, S., Weinberger, K. Q. Resource Aware Person Re-Identification Across Multiple Resolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 8042-8051) (2018)

[19] Si, J., Zhang, H., Li, C. G., Kuen, J., Kong, X., Kot, A. C., Wang, G.: Dual Attention Matching Network for Context-Aware Feature Sequence Based Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5363-5372) (2018)

[20] Suh, Y., Wang, J., Tang, S., Mei, T., Mu Lee, K. Part-Aligned Bilinear Representations for Person Re-Identification. In Proceedings of the European Conference on Computer Vision (pp. 402-419) (2018)

[21] Chen, D., Xu, D., Li, H., Sebe, N., Wang, X. Group Consistent Similarity Learning via Deep CRF for Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 8649-8658) (2018)

[22] Wang, C., Zhang, Q., Huang, C., Liu, W., Wang, X. Mancs: A Multi-task Attentional Network with Curriculum Sampling for Person Re-Identification. In Proceedings of the European Conference on Computer Vision (pp. 365-381) (2018)

[23] Kalayeh, M. M., Basaran, E., Gkmen, M., Kamasak, M. E., Shah, M. Human Semantic Parsing for Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1062-1071) (2018)

[24] Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y. Random Erasing Data Augmentation. arXiv preprint arXiv:1708.04896 (2017)

[25] Huang, H., Yang, W., Chen, X., Zhao, X., Huang, K., Lin, J., Du, D. EANet: Enhancing Alignment for Cross-Domain Person Re-identification. arXiv preprint arXiv:1812.11369 (2018)