Improving Person Re-Identification Performance Using Body Mask Via Cross-Learning Strategy

Junyi Wu[1], Lingxiang Yao[2], Yan Huang[2], Jingsong Xu[2], Qiang Wu[2], and Liqin Huang[1]

[1]College of Physics and Information, Fuzhou University, Fuzhou, China
[2]School of Electrical and Data Engineering, University of Technology Sydney, Sydney, Australia

***Abstract*—The task of person re-identification (re-id) is to find the same pedestrian across non-overlapping cameras. Normally, the performance of person re-id can be affected by background clutters. However, existing segmentation algorithms are hard to obtain perfect foreground person images. To effectively leverage the body (foreground) cue, and in the meantime pay attention to discriminative information in the background (*e.g.*, companion or vehicle), we propose to use a cross-learning strategy to take both foreground and other discriminative information into account. In addition, since currently existing foreground segmentation result always involves noise, we use Label Smoothing Regularization (LSR) to strengthen the generalization capability during our learning process. In experiments, we pick up two state-of-the-art person re-id methods to verify the effectiveness of our proposed cross-learning strategy. Our experiments are carried out on two publicly available person re-id datasets. Obvious performance improvements can be observed on both datasets.**

***Index Terms*—Person re-identification, body segmentation, cross-learning**

Fig. 1. Samples from Market-1501.The first row is the original images, the second row is the segmented body masks, and the last row is the foreground images.

I. Introduction

The task of person re-identification (re-id) attempts to tackle the problem of matching pedestrians across non-overlapping camera views [1]. Person re-id is a very challenging task due to variations of pose, occlusion, illumination, viewpoints, and background. Background interference has been regarded as the main factor that influences the accuracy. Mismatching can be observed for images with a similar background, but belonging to different classes [2], [3].

Focusing on foreground segmentation can directly eliminate the background interference and improve the performance of the person re-id in various background scenarios. [4] proved that body masks were robust to illumination and cloth colors. [5] introduced a neural network using a human parsing module and the random-background data augmentation to address the person re-id problem. The above-mentioned methods depend on a strict assumption that background interference should be totally eliminated from the segmented human masks. However, even using the most advanced segmentation algorithms (*e.g.*, Mask R-CNN [6] or JPPNet [7]), there still exists some noise in the segmented masks. This noise can cause misunderstandings and have a negative impact on person re-id. Fig. 1 presents some segmented masks using Mask R-CNN [6]. It can be seen that there exists some noise in the segmented masks.

Since the background is removed, some discriminative cues, such as a red backpack, are also removed from the foreground images. We think it is inadvisable to neglect the discriminative contents, because sometimes these discriminative contents can serve as useful context cues. In [8], original images were used as the inputs to remedy the incomplete or false segmentation. [9] used synthetic RGB-Mask pairs as input. A mask-guided contrastive attention model was applied to learn features from the body and background regions respectively. However, [8], [9] only used masks to eliminate the background interference, but completely ignored other discriminative information. Our proposed method makes an improvement on this disadvantage.

In this paper, we propose a more comprehensive model for person re-id. A cross-learning strategy is proposed to connect foreground segmentation and the other discriminative contents. Features learned from this strategy are used as complementary cues to help improve the re-id performance. The input of our proposed method and [8] are similar. But during the following procedure, a cross-learning strategy is adopted in our method, which takes foreground information and other discriminative contents into consideration. It is the main difference between our paper and [8], and it is one of the main contributions of this paper. The contributions of this paper are as follows,

1) We propose a new method to connect foreground information with some other discriminative contents by adopting a cross-learning strategy. Features generated from this strategy can be applied as complementary cues for person re-id. Label Smoothing Regularization (LSR) is also utilized to reduce the chances of over-fitting during the model training procedure.

2) We evaluated our proposed method on two widely-used

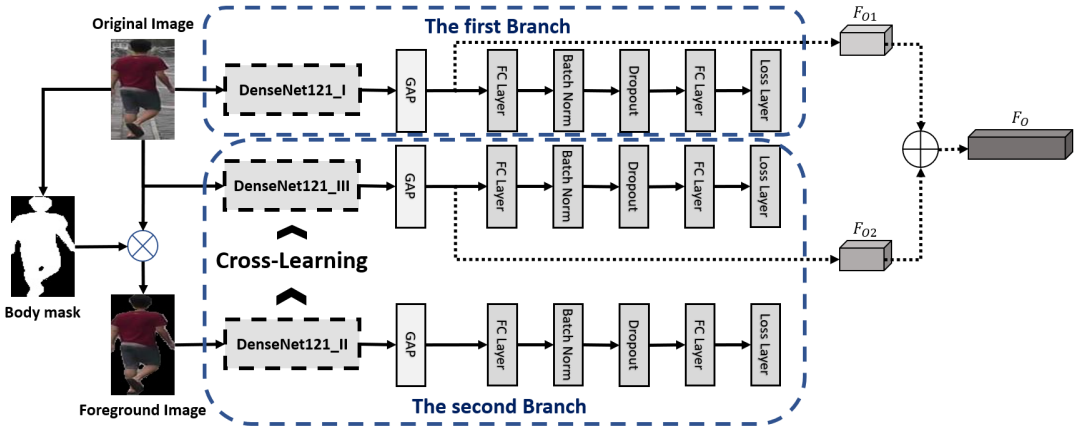

Fig. 2. Architecture of our proposed model. Our model consists of two branches. The proposed cross-learning strategy is mainly used for the second branch to retrain the DenseNet121_III.

and challenging datasets for person re-id, Market-1501 [10] and DukeMTMC-ReID [11]. The effectivity and flexibility of our method have been verified by the related experiments.

II. PROPOSED METHOD

A. Model Architecture

As shown in Fig. 2, our proposed model is composed of two branches, and these two branches both follow the conventional fine-tuning strategy in [12].

Although both branches require the original images as the input, these images are utilized in different stages and for different purposes. The first branch is used to extract the global features, and the original images are fed into DenseNet121_I from the beginning. The second branch uses the cross-learning strategy to combine foreground with some other discriminative contents and to generate a more discriminative representation. The original images join in the second training stage halfway to retrain DenseNet121_III.

For the second branch, a background segmentation method is first used to obtain the foreground masks. The foreground images are extracted from the original images according to the segmented masks. In this paper, we choose JPPNet [7] as the segmentation method, since JPPNet is more invariant to the background changes. The extracted foreground images are mainly employed for the DenseNet121_II fine-tuning.

Taken the original image I_i as input, the procedure for feature extraction can be formulated as follows,

$$\mathcal{F}_{o1} = \mathcal{P}_{GAP}(\mathcal{B}_{DenseNet}\{I_i; W_I|W_{ImageNet}, b_I|b_{ImageNet}\}) \quad (1)$$

$$\mathcal{F}_{o2} = \mathcal{P}_{GAP}(\mathcal{B}_{DenseNet}\{I_i; W_{III}|W_{II}, b_{III}|b_{II}\}) \quad (2)$$

where $\mathcal{F}_{o1}$ and $\mathcal{F}_{o2}$ denote the features from the two branches. $\mathcal{P}_{GAP}$ is the Global Average Pooling. $\mathcal{B}_{DenseNet}$ is the backbone network DenseNet121. $W_I|W_{ImageNet}$ and $b_I|b_{ImageNet}$ respectively denote the weight and the bias of DenseNet121_I based on the prior knowledge $W_{ImageNet}$ and $b_{ImageNet}$ provided by ImageNet. $W_{III}|W_{II}$ and $b_{III}|b_{II}$ respectively denote the weight and the bias of DenseNet121_III based on the prior knowledge W_{II} and b_{II}, which are learned from the foreground images.

The final concatenated feature $\mathcal{F}_o$ can be formulated as,

$$\mathcal{F}_o = \mathcal{F}_{o2} \oplus \mathcal{F}_{o2} \quad (3)$$

where $\oplus$ denotes the concatenation operation.

B. Cross-Learning Strategy

For person re-id, some useful cues might be lost if we only focus on the extracted foreground information and neglect the other discriminative contents. Firstly, it is not sufficient for the foreground knowledge to learn the intrinsic characteristics of a target person. Secondly, some content information can be used as useful context cues in person re-id. Besides, even using the state-of-the-art segmentation methods, some noise still exist in the segmented foreground images.

We apply a cross-learning strategy to handle these problems. On one hand, it can reveal the correlation between foreground and other discriminative cues. On the other hand, it can remedy this foreground information scarcity problem without adding much background inference. Our cross-learning strategy can be separated into two training stages. During the first stage, our model aims at exploring the foreground information from the extracted foreground images. In this stage, our model neglects the intrinsic correlation within the images. In the second stage, our model focuses on learning the discriminative information by using the original images as input. The intrinsic correlation between foreground and other discriminative contents is taken into account in this training stage, which can make up for the neglected correlation in the first learning stage and contribute to providing sufficient complementary cues.

As shown in Fig. 2, the pre-trained DenseNet121_II model is firstly finetuned with the extracted foreground images. After that, the fine-tuned model is retrained with the original images. In this way, our retrained DenseNet121_III model can not only pay attention to the extracted foreground information, but also

take the intrinsic correlation between foreground information and the other useful discriminative contents into consideration. Compared with DenseNet121_II, DenseNet121_III pays more attention to the neglected discriminative contents. Compared with DenseNet121_I, DenseNet121_III associates foreground segmentation and some other discriminative contents in a more effective way. In Section III, we will demonstrate our proposed model with the proposed cross-learning strategy can present a better performance than most re-id methods.

C. Label Smoothing Regularization (LSR)

In this paper, we employ LSR to reduce the chance of overfitting. LSR assigns a smaller value to those non-ground truth classes rather than directly assigns zero to them. By doing so, it can prevent the network from biasing towards the ground truth classes and ignoring the non-ground truth ones. In LSR, $q_{LSR}(k)$ can be formulated as follows,

$$q_{LSR}(k) = \begin{cases} \frac{\epsilon}{k} & k \neq y \\ 1 - \epsilon + \frac{\epsilon}{k} & k = y \end{cases} \tag{4}$$

where $q_{LSR}(k)$ means the ground truth distribution, $\epsilon \in [0, 1]$, $k \in \{1, \quad 2, \quad \cdots\cdots, \quad K\}$, K is the number of classes, and y denotes the ground truth class label. In this way, the cross-entropy loss function can be changed into,

$$l_{LSR\_loss} = (1 - \epsilon) log(p(y)) - \frac{\epsilon}{K} \sum_{k=1}^{K} log(p(k)) \tag{5}$$

where $p(k)$ is the predicted probability of the input belonging to label k.

III. EXPERIMENT

In this section, our proposed method has been evaluated on Market-1501 and DukeMTMC-reID. The cumulative matching characteristics (CMC) at Rank-1 and mean average precision (mAP) is used as the performance evaluation index. Compared with the other results without re-ranking, we can conclude that our method by adopting the cross-learning strategy contributes to improving the performance of person re-id.

A. Experiments on Market-1501

Market-1501 contains 32,668 images of 1501 identities from 6 cameras. The training set contains 12,936 images of 751 identities, and the testing set contains the left of the other 750 identities. We follow the standard setting in [10], and for each identity in the test set, one image from each camera is selected as the query image.

Experiments on Market-1501 can be divided into three parts. For the first experiment, we concentrate on explicating that it is rational and effective for person re-id to take full advantage of the correlation between foreground segmentation and some other discriminative contents by adopting our proposed cross-learning strategy. For the second experiment, we focus on indicating that this learned correlation can be used as complementary cues to improve the accuracy for person re-id.

TABLE I
RESULTS WITH RIM ON MARKET-1501

Model	Rank-1(%)	Rank-5(%)	mAP(%)
RIM	90.7	96.8	75.7
CLM	89.5	96.0	71.1
Fusion	**92.1(+1.4%)**	**97.1**	**77.6(+1.9%)**

TABLE II
RESULTS WITH PCB ON MARKET-1501

Model	Rank-1(%)	Rank-5(%)	mAP(%)
PCB	92.1	97.2	77.0
CLM	89.5	96.0	71.1
Fusion	**93.5(+1.4%)**	**97.4**	**79.9(+2.9%)**

TABLE III
COMPARE WITH OTHER METHODS ON MARKET-1501

Methods	Rank-1(%)	mAP(%)
SVDNet (ICCV17) [13]	82.3	62.1
PDC (ICCV17) [14]	84.1	63.4
TriNet [15]	84.9	69.1
JLML (IJCAI17) [16]	85.1	65.5
MpRL (TIP19) [17]	85.75	67.53
HA-CNN (CVPR18) [18]	91.2	75.7
OURS	**92.1**	**77.6**
OURS+LSR	**92.5**	**78.2**

PCB [19] has been chosen as the representative method in this experiment, because it has been considered as one of the best person re-id approaches. We also compare our method with some state-of-the-art person re-id methods in our last experiments. Table I-VI demonstrate the experiment results. In these tables, RIM denotes the first branch, and CLM denotes the second branch, as mentioned in Section II-A.

Table I indicates the results on Market-1501 for RIM, CLM and their fusion. It can be intuitively seen from this table that the fused model outperforms RIM by 1.4% and CLM by 2.6% in Rank-1. Besides, the mAP of the fused model increases to 77.6%, 1.9% higher than RIM and 6.5% higher than CLM. Rather than simply use foreground information, CLM focuses on exploring the connection between foreground segmentation and other discriminative contents, and contributes to obtaining a better representation of the person intrinsic characteristics.

Table II also presents the results with PCB on Market-1501. We can see that the fusion accuracy increases to 93.5%, 1.4% higher than PCB solely being used. Following the instructions in [19], we retrained PCB on Market-1501. PCB achieves an excellent performance. But with our proposed complementary strategy, the re-id performance still can be improved a lot.

Table III compares our proposed method with some state-of-the-art person re-id methods. Except for our method, the other results are directly quoted from their original papers. Even if without PCB, our method still outperforms other methods.

In all, it is reasonable for person re-id to learn the correlation between foreground information and some other discriminative contents by adopting the cross-learning strategy. The obtained correlation can be used as the complementary cues to improve the person re-id performance.

TABLE IV
RESULTS WITH RIM ON DUKEMTMC-REID

Model	Rank-1(%)	Rank-5(%)	mAP(%)
RIM	82.5	91.4	65.7
CLM	80.8	89.9	61.6
Fusion	**83.9(+1.4%)**	**91.8**	**67.9(+2.2%)**

TABLE V
RESULTS WITH PCB ON DUKEMTMC-REID

Model	Rank-1(%)	Rank-5(%)	mAP(%)
PCB	83.5	91.7	69.4
CLM	80.8	89.9	61.6
Fusion	**85.1(+1.6%)**	**92.2**	**70.7(+1.3%)**

TABLE VI
COMPARE WITH OTHER METHODS ON DUKEMTMC-REID

Methods	Rank-1(%)	mAP(%)
PAN (TCSVT18) [20]	71.6	51.5
SVDNet (ICCV17) [13]	76.7	56.8
MpRL (TIP19) [17]	76.81	58.56
DPFL (TPAMI18) [21]	79.2	60.6
HA-CNN (CVPR18) [18]	80.5	63.8
Deep-Person [22]	80.9	64.8
OURS	**83.9**	**67.9**
OURS+LSR	**84.4**	**68.3**

B. Experiments on DukeMTMC-reID

DukeMTMC-reID contains 36,411 images from 8 cameras. The training set contains 16,522 images of 702 identities, and the testing set contains 702 identities, 2,228 query images and 17,661 gallery images.

Experiments on DukeMTMC-reID can be divided into three parts, as experiments on Market-1501. Table IV-VI illustrate the relevant experiment results. The rank-1 of our fused RIM and CLM increases to 83.9%, 1.4% higher than RIM and 3.1% higher than CLM. The Rank-1 of the fusion of PCB and CLM rises to 85.1% and the fused mAP also rises to 70.7%. Besides, results in Table VI indicate that our proposed method outperforms most state-of-the-art person re-id methods.

In conclusion, Table I-VI have verified the effectiveness of our proposed method. It is easy for the proposed method to be adapted to different databases. Also, relevant experiments on DukeMTMC-reID have proved that the learned correlation between foreground segmentation and some other discriminative contents can be used as cues to enhance the re-id performance.

The purpose of our paper is not to propose a state-of-the-art method, but to propose a useful complementary strategy for person re-id. Our proposed method can be applied by most methods to enhance their performance. Relevant experiments have verified the effectiveness and flexibility of our proposed method. The experiments also indicate that as a complement, features generated from our proposed cross-learning strategy conduce to improving the re-id performance.

IV. CONCLUSION

We propose an universal complementary strategy for person re-id in this paper, and this proposed strategy can be used by most person re-id methods to enhance their performance. This proposed cross-learning strategy aims at revealing the intrinsic correlation between foreground segmentation and some other discriminative contents. Features learned from this correlation take full advantages of the foreground body cues and the other useful contexts. Experiments have been carried out on Market-1501 and DukeMTMC-reID. Relevant results indicate that our proposed method can be applied as a complementary strategy to improve the re-id performance.

V. ACKNOWLEDGMENT

This work was supported by Major Science and Technology Projects in Fujian, China (2018H0018). As the corresponding author, Liqin Huang's email address is hlq@fzu.edu.cn.

REFERENCES

[1] M. Dikmen, E. Akbas, T. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," in *ACCV*, 2010, pp. 501–512.
[2] T. B. Nguyen, V. P. Pham, T.-L. Le, and C. V. Le, "Background removal for improving saliency-based person re-identification," in *KSE*, 2016, pp. 339–344.
[3] Y. Huang, Q. Wu, J. Xu, and Y. Zhong, "Sbsgan: Suppression of inter-domain background shift for person re-identification," in *ICCV*. IEEE, 2019.
[4] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *TPAMI*, vol. 25, no. 12, pp. 1505–1518, 2003.
[5] M. Tian, S. Yi, H. Li, S. Li, X. Zhang, J. Shi, J. Yan, and X. Wang, "Eliminating background-bias for robust person re-identification," in *CVPR*, 2018, pp. 5794–5803.
[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *ICCV*, 2017, pp. 2980–2988.
[7] X. Liang, K. Gong, X. Shen, and L. Lin, "Look into person: Joint body parsing & pose estimation network and a new benchmark," *TPAMI*, 2018.
[8] L. Qi, J. Huo, L. Wang, Y. Shi, and Y. Gao, "Maskreid: A mask based deep ranking neural network for person re-identification," *arXiv preprint arXiv:1804.03864*, 2018.
[9] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *CVPR*, 2018, pp. 1179–1188.
[10] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015, pp. 1116–1124.
[11] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *ECCV*, 2016, pp. 17–35.
[12] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.
[13] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," in *ICCV*, 2017, pp. 3820–3828.
[14] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *ICCV*, 2017, pp. 3980–3989.
[15] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.
[16] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," in *IJCAI*, 2017, pp. 2194–2200.
[17] Y. Huang, J. Xu, Q. Wu, Z. Zheng, Z. Zhang, and J. Zhang, "Multi-pseudo regularized label for generated samples in person re-identification," *TIP*, vol. 28, no. 3, pp. 1391–1403, 2019.
[18] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *CVPR*, 2018, pp. 2285–2294.
[19] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *ECCV*, 2018, pp. 480–496.
[20] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *TCSVT*, 2018.
[21] Y. Chen, X. Zhu, and G. Shaogang, "Person re-identification by deep learning multi-scale representations," in *ICCV*, 2017, pp. 2590–2600.
[22] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu, and Y. Xu, "Deep-person: Learning discriminative deep features for person re-identification," *arXiv preprint arXiv:1711.10658*, 2017.