HLFNet: High-low Frequency Network for Person Re-Identification

Cen Liu, Lijun Guo, and Rong Zhang

***Abstract*—Person re-identification (re-ID) technology has attracted many scholars in the past few years. With the recent developments of deep learning technology, person re-ID has been greatly improved. However, the main chalenge of re-ID is to distinguish the detailed information in different images. Consequently, it is of significant importance to extract fine-grained features in the re-ID tasks. In the present study, a novel method, called the high-low frequency network (HLFNet), is proposed to effectively use the image information of different frequencies and focus on the detailed information between different individual images. In this regard, high frequency and low-frequency information are initially extracted from the original image, and then two backbones are applied to extract the features from the two information branches. Different frequencies of image information complement each other so that a better recognition effect can be achieved. Moreover, a local branch is utilized to extract the distinguishable local features for guiding the global feature branch in the training stage. Finally, only the extracted global feature from the trained network is required in the inference phase of re-ID. Performed experiments demonstrate that the proposed method can significantly enhance the feature representation accuracy and achieve the state-of-the-art performance on diverse benchmarks.**

***Index Terms*—Person re-identification, different frequency information, local and global feature.**

I. Introduction

With the advent of advanced monitoring equipment, multi-camera tracking technology has attracted many scholars in the past few years. The person re-identification (re-ID) has been proposed to detect the same person in multi-cameras, which is the basis of multi-camera tracking analysis. However, considering different environmental parameters, including the light, viewing angle, occlusion and background of each camera, the same person may have different appearances in different cameras, thereby reducing the person re-ID accuracy. With remarkable developments of neural network technology in the past few years, the convolution neural network [9] has been widely applied to extract the appearance features of person images.

Currently, the majority of person re-ID methods use global or local aggregation features to form a combined representation of person images [2], [13], [22], [23]. Specifically, Zheng *et al.* [22] proposed the ID-discriminative embedding (IDE) model, which considers the training process of re-ID as a multiclass classification task, while each identity is considered as a different class. Liu *et al.* [10] introduced an attention mechanism into the global feature extraction. Moreover, Qi *et al.* [12] proposed a new double stream deep neural network based on RGB and gray-scale information. Further investigations showed that this scheme could effectively improve the generalization aspects of re-ID.

Although reasonable results have been achieved through the above-mentioned methods, fine-grained information containing critical data for the recognition tasks has been rarely considered in these methods. In order to resolve this shortcoming, it is intended to propose an effective framework entitled HLFNet for person re-ID tasks. The guide filter [6] is applied to process the original image and then low- and high-frequency information of the original image is extracted. It should be indicated that low- and high-frequency information refers to the smooth image with rich color information and the sharpened image with outstanding details, respectively. The main purpose of this method is to make the model focus on details and the texture information of the image. Since the local information has a crucial impact on the Reid task, it is of significant importance how to use and manage the local information. To this end, a local branch with a group of convolution filters is applied in the training stage. This branch can be regarded as a learning strategy to guide the training of global features. Unlike the training stage, only the obtained features from the global branch are applied in the inference stage. The main contributions of the present study can be summarized as follows:

1) A novel dual-stream person re-ID network is proposed, which covers both low- and high-frequency information. This is a pioneering scheme to use different frequency information of the original image to enhance the performance of the person re-ID model.
2) A guide learning strategy is proposed, which uses a local branch to guide the global feature learning. This strategy can integrate the discriminability of local features into global features, thereby improving the network capabilities in extracting the global features.

Manuscript received April 30, 2021; revised May 21, 2021; accepted May 24, 2021. Date of publication May 27, 2021; date of current version June 17, 2021. This work was supported in part by Zhejiang Provincial Public Welfare Technology Research Project under Grant LGF21F020008 and in the part by the Ningbo Municipal Natural Science Foundation of China under Grant 2018A610057. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Le Lu. *(Corresponding author: Lijun Guo.)*

The authors are with the Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315000, China (e-mail: liucen05@163.com; guolijun@nbu.edu.cn; zhangrong@nbu.edu.cn).

Digital Object Identifier 10.1109/LSP.2021.3084508

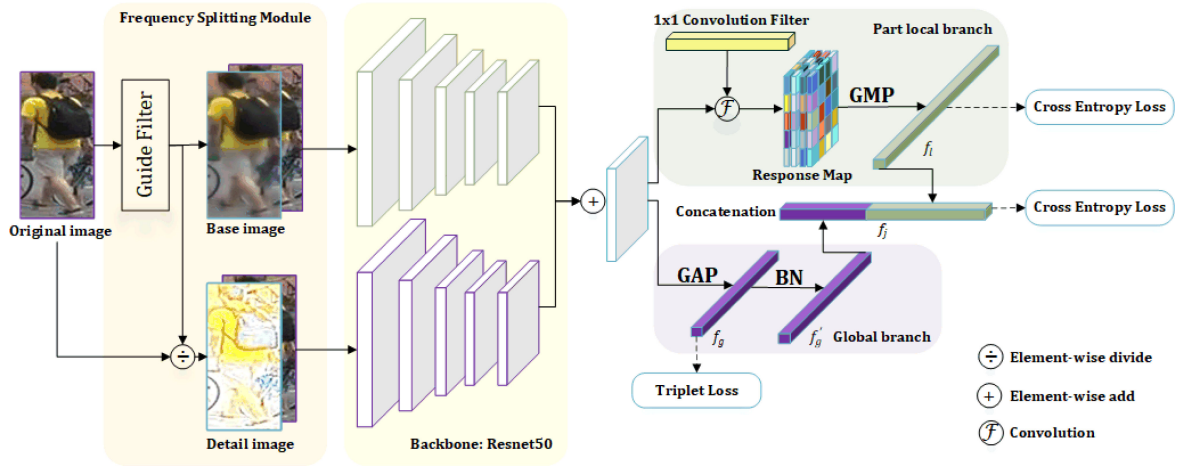

Fig. 1. Architecture of the proposed HLFNet. The frequency-splitting module (FSM) is applied to pre-process the input image and get the base image (i.e. concatenation with the original and low-frequency images) and the detail image (i.e. concatenation with original and high-frequency image). Meanwhile, the local and global branches are applied to process the extracted features from the backbone and then the cross-entropy and triplet losses are utilized to supervise the learning.

3) The proposed method is applied to several standard datasets. Obtained results reveal that through the proposed method, the state-of-the-art performance can be achieved in the person re-ID task.

II. The Proposed Method

A. Network Architecture

As illustrated in Fig. 1, a two-stream network is applied to fuse the original image with the low-frequency and high-frequency image information after processing. Before the input of the backbone, the original image processing should pass the frequency-splitting module (FSM) to separate low- and high-frequency information. This process will be comprehensively described in Section II-B. Then two backbone network branches with the same structure are used to extract feature maps from the base layer and the detail layer, respectively. Inspired by performed investigations [12], the two feature maps are then combined for training.

In the global branch, the global average pooling (GAP) is applied to the previously obtained feature map. In the training stage, a group of 1×1 learnable convolution filter kernel is used as a local branch filter to extract the feature map. Then a response feature map can be obtained after the convolution. The maximum response, which is obtained by the global max-pooling (GMP), is defined as the possible location of the local feature that results in a vector as an output. In this way, the reliability of global feature map extraction can be improved by adding local branches to guide the learning of the whole network. It is worth mentioning that a BNNeck re-ID model module [11] is used in this article to distinguish the triplet lost feature space from the ID loss feature space.

B. Frequency Splitting Module

A key step in the person re-ID task is finding a method to extract more robust and distinguishing features. However, low-frequency information in the original image is more about color and appearance, while high-frequency information mainly contains information about texture and other details. Accordingly, the FSM is applied to transform the input image into the base and detail layer, and then two independent branch networks are used to extract features. This increases image details in the later feature.

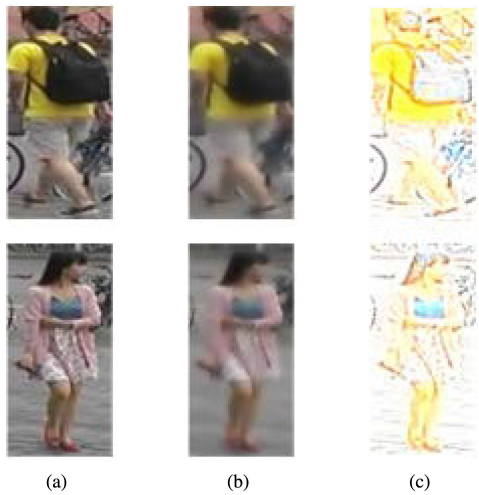

Fig. 2. Examples of images obtained from the FSM: (a) original image, (b) smoothed low-frequency image, and (c) high-frequency image.

Before the input network, the guide filter is applied to decompose the original image I_o into a smooth low-frequency image I_l. Subsequently, the following high-frequency image I_h can be obtained by dividing I_o and I_l:

$$I_l = \mathcal{G}(I_o); I_h = I_o/(I_l + eps) \tag{1}$$

where $\mathcal{G}$ is the guide filter operation and I_o, I_l, I_h denote the original image, blurred low-frequency color image and high-frequency image, respectively. Moreover, eps denote a minimum number to avoid infinity (INF) in the I_h.

Fig. 2 illustrates examples for low-frequency and high-frequency information images obtained from the FSM. It is

observed that I_l contains more color information. On the other hand, I_h contains more edges and texture information, while the color information is scarce. Then I_o is connected to I_l and I_h to get two fused layers, base image and detail layer as the network input.

C. Local Branch Guided Learning

In order to make the global feature more reliable, a local branch is used to constrain the global branch during the training stage. In particular, a 1×1 convolution filter and a GMP is designed in the local branch as a local high-response feature detector. It should be indicated that a 1×1 filter with high response is applied to a certain discrimination region. The size of the feature map is $C \times H \times W$, and the number of filter channels equals the number of corresponding high local regions of image multiplied by the total number of IDs. It is only necessary to select the position of the maximum value in the response feature map while the receptive field can be mapped to the local region of distinguishing features in the original image. The local learnable branch of the feature vector f_l can be expressed as follows:

$$f_l = \mathcal{P}(\mathcal{F}(W^{ck}, X)) \tag{2}$$

where c is the number of ID classes and each class has k corresponding high local regions. In this letter, it is assumed that $k = 5$. Moreover, X denotes the feature map obtained from the backbone. It is necessary to learn a set of 1×1 convolution filter kernels $W^{ck} \in c \times k$. Employing a group of convolution kernels as $\mathcal{F}$, the response map can be obtained while the corresponding channel dimension is ck. Then the GMP in the form of $\mathcal{P}$ is utilized to get the maximum response of each channel and operate on the obtained feature map. Then a full connection (FC) layer and a Softmax layer classifier are used for f_l, while the cross-entropy loss is applied to constrain the branch. The extracted feature f_l from the local branch is concatenated with the extracted feature f'_g from the global branch, and the cross-entropy loss is applied to constrain it.

Therefore, the loss generation depends on all branches. When the gradient back-propagation is performed in the training phase, the global branch undergoes a gradient loss, which can be calculated by the fusion feature f_j from the two branches. It is worth noting that the global branch is always affected by the local branch to adjust network parameters. In other words, the local branch plays a similar regularization role in guiding the learning of features for the global branch in the training phase.

More specifically, the local feature would not be used in the inference stage. The main reasons for this consideration can be summarized as follows: 1) In order to use the local information and based on the number of samples in the whole training set, information extracted of the discriminative location from the local characteristics is used in the training stage to guide the learning process of global features so that the global information has local discriminative performance. 2) Eq. (2) indicates that when this parameter of the local patch number is selected in advance, the dimension of the feature vector f_l is dependent on the total number of training IDs. Meanwhile, when the local features in the training data of different scenes are used in the inference stage, the robustness of the model test is affected, which is unfavorable in practical applications. 3) Only the global branch is applied in the test so that efficiency of the inference stage can significantly improve.

D. Designing the Loss Function

Fig. 1 indicates that the proposed network model has three feature outputs, including feature f_g after the global branch, feature f_l after the local branch, and joint feature f_j concatenated by the local branch f_l and feature f'_g from the global branch after the BN layer. In the training phase, the network is optimized through the following loss function:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{ce}(\sigma_l(W^l f_l), y_i) + \mathcal{L}_{ce}(\sigma_j(W^j[f'_g f_l]), y_i) + \\ \mathcal{L}_{ce}(\sigma_g(W^g f_g\prime, y_i) + \alpha \mathcal{L}_{triplet}(f_g, y_i, y_j)\end{aligned} \tag{3}$$

where $\mathcal{L}_{ce}\{\cdot\}$ and $\mathcal{L}_{triplet}\{\cdot\}$ are the cross-entropy loss and triplet loss, respectively. Moreover, W^l, W^g and W^j denote the FC layer after the f_l feature, f'_g feature and f_j feature, respectively. Meanwhile, σ_j, σ_g and σ_l are the Softmax function in the FC layer of the join feature, global feature and local feature, respectively. Finally, y_i is the training sample, y_j is a negative sample and α is the balanced weight of the triplet loss. The balanced weight parameter is set to $\alpha = 0.1$.

III. Experimental Analysis

A. Datasets and Evaluation Metrics

In the present study, three mainstream person re-ID datasets, including the Market-1501 [21], DukeMTMC [24] and CUHK03 [9] datasets are applied to evaluate the performance of the proposed model. The Market-1501 dataset contains 12,936 images in the training set and 19,732 images in the test set. On the other hand, the DukeMTMC dataset contains 16,522 images in the training set and 17,661 images in the test set. There are 702 people in the training data set. Finally, the CUHK03 dataset contains 14,097 images of 1467 people from 10 cameras. Meanwhile, the mean average precision (mAP) and top-1 accuracy (Rank-1) are mainly applied to evaluate the performance of models.

B. Implementation Details

In all experiments, the ResNet50 dataset [7] pre-trained in ImageNet [8] is selected as the backbone. It is worth mentioning that the number of convolution input channels in the first layer is changed from 3 to 6. The input image size in the training stage is 384×128 and the initial learning rate is 3.5×10^{-4}, and the total training epochs are set to 180. In the experiments, the warm-up strategy [3] is applied. In this regard, the learning rate in the first 10 epochs increases linearly from 3.5×10^{-5} to 3.5×10^{-4}. Conventional data preprocessing strategies in this regard are the horizontal flipping, random clipping and random erasing [26].

C. Comparison With State-of-The-Art Methods

In this section, the performance of the proposed method is compared with that of other advanced person re-ID models on

TABLE I
PERFORMANCE OF DIFFERENT STATE-OF-THE-ART METHODS ON DIFFERENT DATASETS

	Method	Market1501		DukeMTMC		CUHK03			
						Labeled		Detected	
		mAP(%)	Rank-1(%)	mAP(%)	Rank-1(%)	mAP(%)	Rank-1(%)	mAP(%)	Rank-1(%)
Ragid strip based	PCB+RPP(ECCV18)[14]	81.6	93.8	69.2	83.3	-	-	57.5	63.7
	MGN(MM18)[16]	86.9	95.7	78.4	88.7	67.4	68.0	66.0	66.8
Attention based	MHN-6(ICCV19)[1]	85.0	95.1	77.2	89.1	72.4	77.2	65.4	71.7
	BAT-net(ICCV19)[4]	84.7	95.1	77.3	87.7	76.1	78.6	73.2	76.2
	CAMA(CVPR19)[17]	84.5	94.7	72.9	85.8	66.5	70.1	64.2	66.6
	RGA-SC(CVPR20)[19]	88.4	96.1	-	-	77.4	81.1	74.5	79.6
Extra semantic based	P^2-Net(ICCV19)[5]	85.6	95.2	73.1	86.5	73.6	78.3	68.9	74.9
	AANet(CVPR19)[15]	83.4	93.9	74.3	87.7	-	-	-	-
	DSA-reID(CVPR19)[18]	87.6	95.7	74.3	86.2	75.2	78.9	73.1	78.2
	ISP(ECCV20)[27]	88.6	95.3	80.0	89.6	74.1	76.5	71.4	75.2
Ours	**HLFNet**	**88.7**	**96.2**	**80.7**	**89.8**	**79.1**	**82.4**	**76.5**	**79.9**

TABLE II
COMPARISON OF DIFFERENT BACKBONES AND INPUTS ON THE MARKET1501 DATASET

Mehtod		mAP(%)	Rank-1(%)
Backbone	Input		
Single	I_o	84.4	93.9
Double	I_o + I_l	85.0	94.8
	I_o + I_h	85.2	95.6
	I_l + I_h	84.5	94.1
Ours		**88.7**	**96.2**

TABLE III
INFLUENCE OF THE FSM ON THE PERFORMANCE OF SOME STATE-OF-THE-ART METHODS

Method	mAP(%)		Rank-1%	
	Original	+FSM	Original	+FSM
PCB+RPP(ECCV18)[14]	81.6	82.8	93.8	94.5
MGN(MM18)[16]	86.9	87.1	95.7	95.8
BAT-net(ICCV19)[4]	84.7	84.9	95.1	95.4
AANet(CVPR19)[15]	83.4	83.8	93.9	94.4
Ours	**-**	**88.7**	**-**	**96.2**

TABLE IV
COMPARISION OF DIFFERENT PERFORMANCE WITH AND WITHOUT THE LOCAL BRANCH

Method	mAP(%)	Rank-1(%)
w/o Local	87.5	95.1
w Local(Ours)	**88.7**	**96.2**

three datasets. Obtained results are presented in Table I. The performance of four mainstream methods, including the proposed method, rigid strip-based method, attention-based method and extra semantic-based method, are compared. It should be indicated that results are not post-processed with re-rank [25]. It is observed that the proposed model is ahead of the second performance in Market1501, DukeMTMC and CUHK03 datasets by 0.1%, 0.7%, 1.7% and 2.0% in mAP, respectively.

D. Ablation Studies

Comparison of different input images: In Table II, images of the input network are replaced with different inputs to demonstrate the benefits of combining high-frequency and low-frequency information in extracting image features. In this set of ablation study, the following cases are considered: single backbone network with only the original image (I_o), two backbones with original and low-frequency images after the guide filter (I_o + I_l), original image with high-frequency image (I_o + I_h), low-frequency and high-frequency images that are not processed by the guide filter (I_l + I_h). In the latest case, which is considered as the base image and detail input image in the present letter, the original image does not processed by the guide filter to input. It is found that the accuracy in extracting high-frequency and low-frequency information obviously outperforms that of using the original image for single extraction.

In order to evaluate the effectiveness of the proposed FSM, this module is applied in several state-of-the-art methods. In this regard, the FSM is applied to transform the backbone in the feature extraction phase of the original model into two branches. Table III presents the obtained results, indicating that the network accuracy improves after applying the FSM.

Benefit of local branch: In order to measure the capability of the proposed local branch, the performance of the model test results with and without local branches in the training stage are compared. Obtained results in Table IV reveal that on the Market1501 dataset, with the local branch method improves mAP and rank-1 indicators by 1.2% and 1.1%, respectively.

IV. CONCLUSION

In this letter, the HLFNet is proposed to combine the high- and low-frequency information of the original image and use the local branch guide in the global branch training for person re-ID. Conducted experiments demonstrate that the proposed network outperforms other state-of-the-art methods on several benchmark datasets.

REFERENCES

[1] B. Chen, W. Deng, and J. Hu,"Mixed high-order attention network for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 371–381.

[2] G. Chen, C. Lin, L. Ren, J. Lu, and J. Zhou, "Self-critical attention learning for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9637–9646.

[3] X. Fan, W. Jiang, H. Luo, and M. Fei, "Spherereid: Deep hypersphere manifold embedding for person re-identification," *J. Vis. Commun. Image Representation*, vol. 60, pp. 51–58, 2019.

[4] P. Fang, J. Zhou, S. K. Roy, L. Petersson, and M. Harandi, "Bilinear attention networks for person retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8030–8039, 2019.

[5] J. Guo, Y. Yuan, L. Huang, C. Zhang, J.-G. Yao, and K. Han, "Beyond human parts: Dual part-aligned representations for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3642–3651.

[6] K. He, J. Sun, and X. Tang, "Guided image filtering," in Proc. *Eur. Conf. Comput. Vis.*, 2010, pp. 1–14.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[9] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 152–159.

[10] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017.

[11] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019.

[12] L. Qi, L. Wang, J. Huo, Y. Shi, and Y. Gao, "Greyreid: A two-stream deep framework with RGB-grey information for person re-identification," 2019, *arXiv:1908.05142*.

[13] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1179–1188.

[14] 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 *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 480–496.

[15] C.-P. Tay, S. Roy, and K.-H. Yap, "Aanet: Attribute attention network for person re-identifications," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7134–7143.

[16] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 274–282.

[17] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang, "Towards rich feature discovery with class activation maps augmentation for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1389–1398.

[18] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Densely semantically aligned person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 667–676.

[19] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation aware global attention for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3186–3195.

[20] H. Zhao *et al.*, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1077–1085.

[21] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1116–1124.

[22] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv:1610.02984*.

[23] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2138–2147.

[24] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline invitro," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3754–3762.

[25] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1318–1327.

[26] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in Proc. *AAAI*, 2020, pp. 13001–13008.

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

[28] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose-invariant embedding for deep person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4500–4509, Sep. 2019.