

Learning to Disentangle Scenes for Person Re-identification[★]

Xianghao Zang^a, Ge Li^a, Wei Gao^a and Xiujun Shu^b

^aSchool of Electronic and Computer Engineering, Peking University, Shenzhen 518055, China.

^bPeng Cheng Laboratory, Shenzhen 518034, China.

ARTICLE INFO

Keywords:

person re-identification
divide-and-conquer
multi-branch network

ABSTRACT

There are many challenging problems in the person re-identification (ReID) task, such as the occlusion and scale variation. Existing works usually tried to solve them by employing a one-branch network. This one-branch network needs to be robust to various challenging problems, which makes this network overburdened. This paper proposes to **divide-and-conquer the ReID task**. For this purpose, we employ several **self-supervision operations** to simulate different challenging problems and handle each challenging problem using different networks. Concretely, we use the **random erasing operation** and propose a novel **random scaling operation** to generate new images with controllable characteristics. A general multi-branch network, including one master branch and two servant branches, is introduced to handle different scenes. These branches learn collaboratively and achieve different perceptive abilities. In this way, the complex scenes in the ReID task are effectively disentangled, and the burden of each branch is relieved. The results from extensive experiments demonstrate that the proposed method achieves state-of-the-art performances on three ReID benchmarks and two occluded ReID benchmarks. Ablation study also shows that the proposed scheme and operations significantly improve the performance in various scenes.

1. Introduction

Person re-identification (ReID) has drawn increasing attention in computer vision society. Given a person image from the query, ReID aims to find all images of the same person from the gallery. In practice, the ReID task has widespread applications in social security and surveillance systems. For example, with the help of surveillance cameras, it can help find out the suspect criminals, look for a lost child in a large mall, etc [70].

Despite achieving much progress [38] [74], it is still challenging to handle various complex scenes in the ReID task, such as scale variation, occlusion, false detection, and a similar appearance. Most existing approaches can be categorized as the one-branch network. And the challenging problems overburden these one-branch networks. To achieve good performance, these one-branch networks have to utilize sophisticated designs or employ additional information (semantics, pose information, *etc.* [28] [40] [55] [49]). These elaborate designs make the one-branch network complicated and over-engineered.

Recently, multi-branch networks have shown their potentials in many fields of deep learning, such as image classification [73] [65], knowledge distillation [51] [1], cross-domain/modality learning [16] [35]. For the ReID task,

many multi-branch networks were also proposed. These networks fall into two categories. The first category borrows thoughts from **knowledge distillation** [45] [80]. They usually have two networks, teacher and student, and use a two-stage training process. The teacher is trained for complex tasks in the first stage. In the second stage, the teacher network is fixed, and its knowledge is transferred to the student. Although this method has two branches, only one branch learns in each training stage. Therefore, it is still under the paradigm of a one-branch network. The second category employs two equal networks and let them **co-teach** [61] [60] [16]. Their branches share the same responsibilities and support each other, which produces better performance than their one-branch counterpart. However, each branch still needs to deal with various challenging scenes, and its responsibility has not reduced, resulting in limited performance improvement.

To effectively disentangle the complex scenes, we propose a **divide-and-conquer strategy** for the ReID task. We mainly analyze two challenging scenes, *i.e.*, **occlusion** and **scale variation**, as illustrated in Fig. 1. We conquer them one by one and improve the overall performance for the ReID task. To this end, we apply two **self-supervision operations** to the input image to obtain new images with the characteristics of challenging scenes. Concretely, we employ the **random erasing** to generate the occluded scenes and propose **random scaling** to generate the scale variation scenes. In this way, the ReID task is divided into simpler ones. The original image is also kept as the general scenes to provide the missing information for other generated images. We also introduce a new input manner, *i.e.*, **homologous input**. This manner solves the input image misalignment problem and further improves the performance. To conquer each challenging scene, we propose a **multi-branch network**, as illustrated in Fig. 2. There are one **master branch** and two

[★]This work was supported by the Key-Area Research and Development Program of Guangdong Province (2019B121204008), the National Natural Science Foundation of China (61801303 and 62031013), the Guangdong Basic and Applied Basic Research Foundation (2019A1515012031), the Shenzhen Science and Technology Plan Basic Research Project (JCYJ20190808161805519), and the Shenzhen Fundamental Research Program (GXWD20201231165807007-20200806163656003).

*Corresponding author: Wei Gao.

✉ zangxh@pku.edu.cn (X. Zang);

geli@ece.pku.edu.cn (G. Li); gaowei262@pku.edu.cn (W. Gao); shuxj@pcl.ac.cn (X. Shu)

ORCID(s):

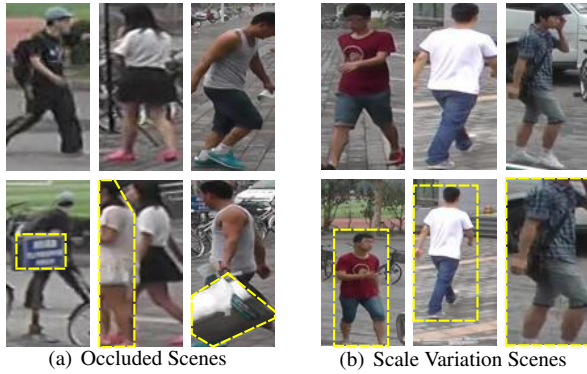


Figure 1: Two typical challenging scenes in ReID. The top and bottom rows show normal and difficult scenes. (a) Occluded scenes. The occluded areas are within the yellow dotted lines. (b) Scale variation scenes. The sizes of pedestrians in the yellow boxes change.

servant branches in this framework. Each servant branch is assigned to deal with one specific challenge. And the master branch is designed to handle the general scenes. The traditional one-branch network needs to deal with occlusion, scale variation, *etc.* In our framework, each servant branch only needs to deal with occlusion or scale variation. Therefore, the burden of each servant branch is relieved. To train the multi-branch network, we employ **mutual learning** to transfer the knowledge between different branches. These branches learn collaboratively and promote each other. For the master branch, we use the original input image without any artificial change. The knowledge from the servant branches benefits the master branch and decreases its overfitting possibility for general scenes. For the servant branch, the artificial image loses some information due to the self-supervision operations. The knowledge from the master branch makes the servant branch implicitly learn the missing information and obtain robustness for a specific scene. In the testing process, features from multiple branches are concatenated as the whole feature representation. Since each branch has a different perceptive ability, the concatenated feature is robust for various scenes.

We evaluate the proposed scheme on three ReID benchmarks, including Market1501 [69], DukeMTMC-reID [47], MSMT17 [57], and two large-scale occluded ReID benchmark, P-DukeMTMC-reID [79], and Occluded-DukeMTMC [40]. The experiments demonstrate our method achieves state-of-the-art performances. An extensive ablation study also shows that the proposed scheme improves the robustness in various scenes.

The main contributions of this paper can be summarized as follows:

- For the challenging ReID task, we introduce a divide-and-conquer strategy to deal with it, which effectively reduces the network learning burden.
- Unlike the traditional multi-branch networks, knowledge communication from different scenes improves

the overall performance, which capitalizes on the potentials of multi-branch networks.

- Experiments on three ReID benchmarks and two occluded ReID benchmarks show that our scheme achieves state-of-the-art performances. The ablation study also demonstrates the effectiveness of the proposed scheme in various scenes.

The rest of this paper is organized as follows. The related works are reviewed and discussed in Section 2, and then we elaborate on the proposed method in Section 3. Experimental results and analysis are presented in Section 4, and finally, Section 5 concludes this paper.

2. Related Work

2.1. Person Re-identification

For the ReID task, the problem of misalignment introduced by scale variation and occlusion has aroused great interest in the computer vision community. Many works explored this problem [18] [19] [26] [37] [13]. Luo *et.al* [37] proposed a Dynamically Matching Local Information (DMLI) to align the local information dynamically. The DMLI calculates the distances between different parts of possible image pairs through a dynamic programming strategy. Miao *et.al* [40] employed the pose estimator to generate landmarks. These landmarks are utilized to indicate the model to focus on the non-occluded regions to overcome the noise introduced by the various obstacles. In this way, they obtained aligned feature representations for the ReID task.

The methods above explicitly achieved the feature alignment with the help of various supporting information. Others deal with this problem in a implicit manner. Zhou *et.al* [76] proposed Omni-Scale Networks (OSNet), which introduced an aggregating gate to fuse feature from different scales to achieve an omni-scale feature representation. The OSNet handles the misalignment by aggregating the multi-scale features and achieves good performance. Jin *et.al* [28] proposed Semantics Aligning Network (SAN), which employed a decoder to reconstruct a dense semantics aligned full texture image. They supervise the ReID task and the semantic texture generating process simultaneously to learn a semantics-aligned feature representation. Quan *et.al* [46] employed the Neural Architecture Search (NAS) to find a part-aware network in a retrieval-based search space automatically.

These methods above can be categorized as a one-branch network. These one-branch networks need to be robust to various challenging problems in the ReID task, which makes these networks overburdened. Although these methods utilized sophisticated designs or additional information to improve the performance, these endeavors make the one-branch network complicated and bring a limited improvement.

2.2. Multi-Branch Networks

There are many multi-branch networks for the ReID task [80] [41] [60] [27] [16] [66]. The first category uses a teacher

network to teach a student network. Porrello *et.al* [45] introduced multiple Views Knowledge Distillation (VKD), which trains the teacher network using multiple views and only gives the student a small set of input views. After the knowledge distillation process, the student outperforms his teacher in the image-to-video setting. Zhuo *et.al* [80] employed a teacher-student framework for occluded ReID. They train the teacher network with a co-saliency network to simulate the occluded ReID, which enables the teacher to perceive the occlusion. Then they use the teacher network to generate the occluded mask to supervise the student network. These methods above employed a two-stage training process where only one network is trained in each stage. Therefore, they still under the paradigm of a one-branch network.

The second category trains each branch simultaneously and makes them co-teach. Yang *et.al* [60] proposed asymmetric co-teaching for the cross-domain ReID. They employed two branches and fed them with samples as pure as possible and as miscellaneous as possible, respectively. To achieve this goal, they encouraged their two networks to promote each other. Ge *et.al* [16] proposed Mutual Mean-Teaching (MMT) for the cross-domain ReID. They employed mean network, soft classification loss, and soft triplet loss to let two networks mutual-teach. The mean network is updated using the running average mean weight of each network. Zhang *et.al* [65] proposed Deep Mutual Learning (DML) and gave two branches the same optimization objective. Although using one branch can complete this task, the DML scheme can find a much wider minimum for its loss function and provide a better generalization performance. However, each branch in these methods still needs to deal with various challenges, resulting in a limited performance improvement.

2.3. Person Re-identification in A Specific Scene

There are various challenging scenes in the ReID task. However, there are no benchmarks designed for the scale variation scene. On the other side, two large-scale benchmarks, P-DukeMTMC-reID and Occluded-DukeMTMC, are proposed for the occluded scenes recently. The occluded ReID has raised increasing attention from the computer vision community [21] [15] [40] [49] [13] [14]. In this field, Miao *et.al* [40] introduced Pose-Guided Feature Alignment (PGFA) and exploited pose landmarks to disentangle the useful information from the occlusion noise. However, this method largely depends on an accurate human pose estimator to detect human landmarks. Sun *et.al* [49] introduced Visibility-aware Part Model (VPM) and employed self-supervision learning to enable the model visibility-aware. Due to the limited self-supervision, the VPM learns a coarse division strategy, which limits its performance.

These methods above merely focused on a specific scene and may fail to handle the ReID task in general scenes. On the contrary, our scheme is effective in various scenes. We divide the challenging scenes in the ReID task into multiple simpler ones and conquer them individually. Each branch achieves the perceptive ability for a particular scene in the

training process. Concatenating features from each branch aggregate these different perceptive abilities and produce significant performance improvement.

3. Learning to Disentangle Scenes

This section elaborates on the proposed method for ReID, *i.e.*, Learning to Disentangle Scenes (LDS). The framework of the proposed LDS method is illustrated in Fig. 2. In this framework, we adopt the design philosophy of **divide-and-conquer** to deal with the ReID task.

3.1. “Divide” the Complex Scenes

This paper identifies the occlusion and scale variation scenes from the complex scenes in the ReID task. We adopt a self-supervision operation to generate new images with the controlled characteristics. First, we apply a *random data augmentation* strategy to each image. The *random data augmentation* introduces more samples for the network training. Then we make three copies from the new samples.

Occlusion Scenes. To generate an image with occlusion, we apply the *random erasing* to the first copy. The probability of *random erasing* is set to 1 to ensure occlusion exists in this image.

Scale Variation Scenes. To generate an image with scale variation, we propose the *random scaling* and apply it to the second copy. The *random scaling* is described in detail below. We first generate a baseboard with the mean value of three channels (R, G, B) of all the images in ImageNet. Then we scale the second copy to $0.8 \sim 1.1$ times its original size. The zoom value is randomly generated. If the zoom value is less than 0.9, the scaled image is pasted in the baseboard center. For the ReID task, the center of input images is often informative, and the marginal part usually contains background and noise information. Putting it at the center of the baseboard makes the servant branch focus on the center of input images. If the zoom value is between 0.9 and 1.0, the scaled image is pasted anywhere on the baseboard. If the zoom value is more than 1.0, the scaled image is pasted at the baseboard center. All marginal parts beyond the baseboard boundary are discarded. The probability of the *random scaling* operation is also set to 1. We set the minimum zoom value to 0.8 because a smaller margin around the image can improve the performance. Meanwhile, a much larger zoomed image introduces more information loss. Through a co-teach strategy in Section 3.3, the master and servant branches focus on the image center, making each branch neglect the marginal part and improve the overall performance.

General Scenes. We keep the third copy without any artificial change. The reason is explained below. The first and second copies are manipulated by the *random erasing* and the *random scaling*, respectively. Thus some useful information in them is lost. We keep the third copy as the original one to provide the missing information to the first and

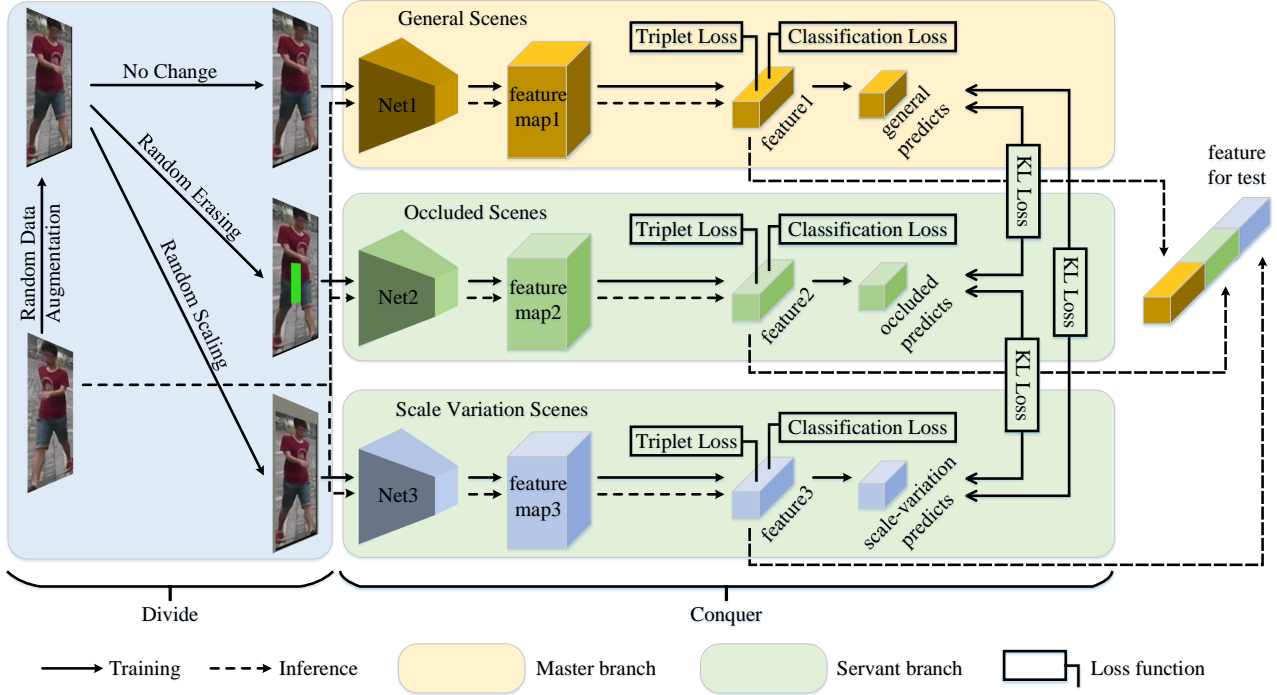


Figure 2: The framework of the proposed LDS method. We propose a divide-and-conquer strategy to deal with the ReID task. This framework contains two parts, “divide” and “conquer” parts. They are used to “divide” the complex scenes and “conquer” the specific scene. In the “conquer” part, we use mutual learning to promote each branch.

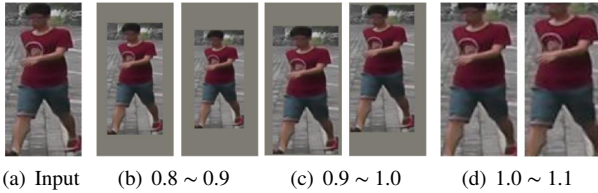


Figure 3: Random scaling. (a) Original input image. (b), (c), (d) Examples when the zoom scale in different ranges.

second ones. On the other side, the third copy represents the images that happened in general scenes.

Image Alignment for Different Scenes. There are misalignment problems for the traditional multi-branch networks, as illustrate in Fig. 4(a). This misalignment problem is mainly derived from the different input images, denoted as the *heterologous input*. In general, the *random data augmentation* operation usually includes *random flipping* and *random cropping*. These operations make the input images change their orientations and center positions, which makes the different branches receive misaligned images. Therefore, the *heterologous input* utilized by most multi-branch networks often results in a misalignment problem. We propose *homologous input* to solve this problem, as illustrated in Fig. 4(b). The self-supervision operation is after the *random data augmentation*. This fashion ensures that the different branches have the same source image. This *homologous input* is simple but effective, and the latter extensive

ablation studies demonstrate its effectiveness.

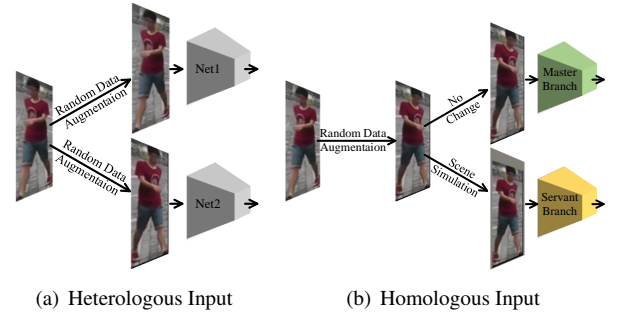


Figure 4: Illustrations of *heterologous input* and the proposed *homologous input*. The *homologous input* is applied to our scheme in Fig. 2.

3.2. “Conquer” the Specific Scene

We propose a multi-branch network to “conquer” each specific scene. This network consists of three branches, including the master branch and two servant branches. The master branch deals with the ReID problem in general scenes, while servant branches handle the occluded scenes and scale variation scenes. We formulate this optimization process for each branch. Given N image samples $\mathcal{I} = \{x_i\}_{i=1}^N$, there are corresponding person IDs as $\mathcal{Y} = \{y_i\}_{i=1}^N$ where $y_i \in \{1, 2, \dots, M\}$. M is the total number of person identities. After the process of *homologous input*, the input image I_i becomes three copies with different characteristics de-

noted as $I_i^{general}$, $I_i^{occlude}$, and I_i^{scale} , which are fed to the master branch, the servant branch for the occluded scenes, and the servant branch for the scale variation scenes, respectively. We employ Θ to represent each network and use I_i^Θ to represent the generated image, $I_i^{general}$, $I_i^{occlude}$, and I_i^{scale} , for each branch. The feature map extracted from one specific branch is denoted as F_i^Θ . Then, each feature map is processed by average pooling and BN Neck layer to generate a normalized feature f_i^Θ .

The first loss function for optimizing each branch is the **triplet loss**. We employ existing **soft margin triplet loss with batch hard mining** [23], which is calculated as follows:

$$\mathcal{L}_{triplet}^\Theta = \frac{1}{B} \sum_{j=1}^B \sum_{a \in b_i} \ln \{1 + \exp [\xi + \max_{p \in \mathcal{P}(a)} d(f_a^\Theta, f_p^\Theta) - \min_{n \in \mathcal{N}(a)} d(f_a^\Theta, f_n^\Theta)]\}, \quad (1)$$

where b_i is the i^{th} batch, B is the number of batches in each benchmark, a, p, n are anchor, positive, and negative samples, respectively, $\mathcal{P}(a)$ and $\mathcal{N}(a)$ are the positive and negative sample sets corresponding to the given anchor a in this batch, ξ is the distance margin threshold, and function $d(\bullet)$ calculates the Euclidean distance between two extracted features. In Eq. 1, we follow the previous research [9] [64] [23] and replace the hinge function $[m + \bullet]_+$ with the soft-plus function $\ln(1 + \exp(\bullet))$. The soft-plus function is a smooth approximation of the hinge function and decays exponentially without having a hard cut-off, resulting in a numerically stable implementation.

The second loss function for optimizing each branch is the **classification loss**. We use the **additive margin softmax** (AM-softmax) [52] to calculate this loss as follows:

$$\mathcal{L}_{cls}^\Theta = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{[\gamma(W_{y_i} f_i^\Theta - m)]}}{e^{[\gamma(W_{y_i} f_i^\Theta - m)]} + \sum_{j=1, j \neq y_i}^M e^{[\gamma(W_j f_i^\Theta)]}}, \quad (2)$$

where W_{y_i} and W_j are the weight vectors associated with class y_i and j in the final classification layer, γ is the scaling factor, and m is the margin to distinguish the similarity distance. In Eq. 2, we follow the [31] [32] [53] [52] to calculate the inner product using normalized weights vectors, W_{y_i} and W_j , and normalized feature f_i . The scaling factor γ is set to a constant instead of a learned weight to accelerate the training process.

For each specific scene, the loss function $\mathcal{L}_{conquer}^\Theta$ is formulated as follows:

$$\mathcal{L}_{conquer}^\Theta = \mathcal{L}_{triplet}^\Theta + \mathcal{L}_{cls}^\Theta. \quad (3)$$

3.3. Mutual Learning

We employ mutual learning to make each branch communicate its knowledge to others. ~~Since all the branches have the same source image, their logits should have a similar data distribution.~~ The difference between their logits is mainly derived from the missing information introduced

by the self-supervision operations. Through mutual learning, the servant branch becomes “sensitive” to the specific scene and can “guess” the missing information under the guidance of knowledge from the master branch. In this way, the servant branch gets the perceptive ability for the specific scene. Meanwhile, the master branch receives knowledge from more scenes, which reduces its overfitting probability and improves its robustness.

We use the **Kullback Leibler (KL) Divergence** to quantify the similarity of logits from different branches. We first calculate the probability p^Θ of class m for pedestrian image sample x_i as follows:

$$p_m^\Theta(x_i) = \frac{\exp(\gamma W_m f_i^\Theta)}{\sum_{k=1}^M \exp(\gamma W_k f_i^\Theta)}, \quad (4)$$

where $\gamma W_m f_i^\Theta$ is the logit fed to the “softmax” layer in the branch Θ . We optimize each branch by employing a KL loss which is calculated as follows:

$$\mathcal{L}_{ML}^\Theta = \frac{1}{S-1} \sum_{s=1, s \neq \Theta}^S \sum_{i=1}^N \sum_{m=1}^M p_m^\Theta(x_i) \log \frac{p_m^\Theta(x_i)}{p_m^s(x_i)}, \quad (5)$$

where S is the branch number. The KL loss \mathcal{L}_{KL}^Θ makes the logits from different branches as similar as possible. For each branch, the loss function \mathcal{L}^Θ is formulated as follows:

$$\mathcal{L}^\Theta = \mathcal{L}_{ML}^\Theta + \mathcal{L}_{conquer}^\Theta. \quad (6)$$

In this paper, we train the proposed LDS in an end-to-end manner. There are three losses for all branches, *i.e.*, $\mathcal{L}^{general}$, $\mathcal{L}^{occlude}$ and \mathcal{L}^{scale} . The overall optimization function for our scheme is calculated as follows,

$$\mathcal{L} = \mathcal{L}^{general} + \mathcal{L}^{occlude} + \mathcal{L}^{scale}. \quad (7)$$

3.4. Advantages of Proposed LDS Method

Existing works usually use a one-branch network for the challenging ReID task. There are many challenging problems that need to be dealt with, making the one-branch network overburdened. Many sophisticated designs and additional information are utilized to strengthen the one-branch network. However, these designs make the one-branch network complicated, and the performance improvement is limited.

The Knowledge Distillation (KD) [24] is proposed to enable a small network to become strong through learning the knowledge from a large one. The difficult learning task is assigned to the teacher network. Then the teacher is fixed, and the knowledge is distilled to the student. Although having fewer parameters, the student becomes as strong as the teacher. However, only one network learns in each training stage. Thus these KD methods are still under the paradigm of a one-branch network.

Mutual learning or co-teaching methods, such as DML [65], MMT [16], make two same networks share the responsibility and promote each other. The performance improvement is mainly due to a different network weight initialization. This operation makes each branch learn in different

directions, and the KD loss makes them have an intermediate and better optimization direction. The multi-branch network often achieves better performance than a one-branch network. However, the number of challenging problems is not reduced, and each branch in these methods has the same responsibility, leading to difficulty in improving performance.

The proposed LDS introduces a divide-and-conquer strategy for the ReID task. Through self-supervision operations, the challenges are divided into simpler ones. Each servant branch only needs to deal with a specific challenge, which reduces the burden of each branch. By employing mutual learning, the servant branches can receive the knowledge from the master branch and obtain the ability to recover the missing information. The knowledge from the servant branches contains incomplete information, which reduces the overfitting possibility for the master branch. Therefore, the performance improvement of the proposed LDS is from the knowledge communication of different scenes. Correspondingly, the performance improvement of the traditional co-teaching methods is from the different initial conditions, *i.e.*, random initialization of network weights.

On the other side, the proposed LDS is a general and flexible framework. More image transformation operations can be introduced to deal with other challenging issues, *i.e.*, lighting variation, similar appearance, etc. This paper focuses on the divide-and-conquer strategy, and two scenes can illustrate the effects of this strategy. Therefore, we only investigate two typical scenes, occlusion and scale variation. Experiment results in the latter section demonstrate that the proposed LDS is more effective than the existing one-branch and multi-branch networks.

4. Experiments

4.1. Benchmarks and Evaluation Metrics

We evaluate the proposed LDS on three image-based ReID benchmarks, including Market1501 [69], DukeMTMC-reID [71], MSMT17 [57], and two occluded ReID benchmarks, P-DukeMTMC-reID [79] and Occluded-DukeMTMC [40].

The Market1501 contains 1501 person identities captured by six different cameras on campus. In the training set, 12936 images for 751 persons are used. There are 19732 and 3368 images of the rest 750 person identities for gallery and query in the testing set.

The DukeMTMC-reID benchmark consists of 1812 person identities collected by eight synchronized cameras from campus. There are 16522 images of 702 identities in the training set. There are 17661 and 2228 images of the other 702 identities for gallery and query in the test set.

The MSMT17 contains 4101 person identities captured by a 15-camera network, including 12 outdoor and 3 indoor cameras on campus. The training set includes 32621 images of 1041 identities. There are 82161 and 11659 images of the other 3060 identities for the gallery and query in the test set.

In P-DukeMTMC-reID, there are 12927 images of 665 person identities for training, including 2647 images with

occlusion and 10280 images without occlusion. There are 11216 images from 634 person identities for test, including 2163 images with occlusion for query and 9053 images without occlusion forming the gallery.

The Occluded-DukeMTMC contains 15618 images of 702 person identities in the training set, 17661 images of 1110 person identities in the gallery, and 2210 images of 519 person identities in the query.

The Cumulative Matching Characteristics (CMC) [17] and mean Average Precision (mAP) [69] are reported. We use the Rank- k scores to represent the CMC curve. All the experiments are performed in a single query setting.

4.2. Implementation Details

We use the PyTorch toolbox, FastReID [22], to achieve the proposed LDS. Additionally, we use the ResNet-ibn [20] [42] as our backbone and initialize it by the ImageNet [11] pre-trained model. The non-local layer [56] is also employed in our backbone. Each person image is resized to 384×128 . We set the batch size to 64 and use Adam [29] with initialized learning rate 3.5×10^{-4} to train each benchmark for 60 epochs. We use the cosine annealing part of the SGDR [34] to adjust the learning rate. We also freeze the backbone in the first 2000 iterations for each benchmark to train the network. Then we train the whole multi-branch network for the rest iterations.

4.3. Comparison with State-of-the-Arts

Table 1 represents the performance comparisons between the proposed LDS and other state-of-the-art methods on three popular benchmarks in terms of CMC accuracy and mAP scores. These methods are within two years and include nine attention-based methods, six semantics-based methods, four stripe/part-related methods, three multi-branch networks, and nine other kinds of methods. They are all trained on the standard training sets without depending on additional images or labels. Early literature often used the re-rank [72] technique. This technique can effectively adjust the order of image candidates and improve the mAP scores. In Table 1, we also present the performance of the proposed LDS with the re-rank.

Performances on Market1501. In table 1, compared to the other state-of-the-art methods, the proposed LDS achieves competitive results. There are other three multi-branch methods, PTL [63], CAMA [62], and HBFP-Net [33]. These methods employed the feature map from different layers to form a rich feature representation. Meanwhile, our method gives each branch a different perceptive ability by feeding them images with different characteristics. Thus, the proposed LDS is more easily implemented. And LDS also achieves a better performance than them. The proposed LDS with re-rank also achieves better performance. These extensive comparisons demonstrate the effectiveness of our scheme.

Performances on DukeMTMC-reID. Table 1 shows that the proposed LDS achieves state-of-the-art performance. Com-

Table 1

Performance comparisons with state-of-the-art methods on Market1501, DukeMTMC-reID, and MSMT17.

	Method	Publication	Market1501		DukeMTMC-reID		MSMT17	
			R1	mAP	R1	mAP	R1	mAP
Attention-based	BAT [12]	ICCV19	95.10	87.40	87.70	77.30	79.50	56.80
	ABD-Net [7]	ICCV19	95.60	88.28	89.00	78.59	82.30	60.80
	CAR [77]	ICCV19	96.10	84.70	86.30	73.10		
	SCAL (spatial) [5]	ICCV19	95.40	88.90	89.00	79.60		
	SONA ²⁺³ -Net _{μ} [58]	ICCV19	95.58	88.83	89.38	78.23		
	MHN-6 (PCB) [4]	ICCV19	95.10	85.00	89.10	77.20		
	IANet [25]	CVPR19	94.40	83.10	87.10	73.40	75.50	46.80
	SCSN (3 stage) [8]	CVPR20	95.70	88.50	90.10	79.00	83.00	58.00
Semantics-based	RGA-SC [68]	CVPR20	96.10	88.40			80.30	57.50
	P ² -Net (+triplet loss) [18]	ICCV19	95.20	85.60	86.50	73.10		
	DSA-reID [67]	CVPR19	95.70	87.60	86.20	74.30		
	SAN [28]	AAAI20	96.10	88.00	87.90	75.50	79.20	55.70
	DLBC [6]	ACM MM20	94.60	87.40	88.70	78.50	78.20	55.60
Stripe/Part-related	ISP [78]	ECCV20	95.30	88.60	89.60	80.00		
	Auto-ReID [46]	ICCV19	94.50	85.10			78.20	52.50
	BDB + Cut [10]	ICCV19	95.30	86.70	89.00	76.00		
	RRID [43]	AAAI20	95.20	88.90	89.70	78.60		
Others	HAA [59]	ACM MM20	95.80	89.50	89.00	80.40		
	SFT [36]	ICCV19	93.40	82.70	86.90	73.20	73.60	47.60
	DCDS [2]	ICCV19	94.81	85.80	87.50	75.50		
	VCFL [30]	ICCV19	89.25	74.48				
	MVP Loss [48]	ICCV19	91.40	80.50	83.40	70.00	71.30	46.30
	OSNet [76]	ICCV19	94.80	84.90	88.60	73.50	78.70	52.90
	DSFL [44]	ACM MM20	96.20	89.90	90.20	81.10	84.20	60.70
	NEWTN [39]	NeurIPS20	95.60	89.40			71.50	53.10
multi-branch	M ³ + HA-CNN [75]	CVPR20	96.50	85.20	87.10	72.20	74.30	43.80
	CtF [54]	ECCV20	93.70	84.90	87.60	74.80		
	DML [65]	CVPR18	89.34	70.51				
	PTL + MGN [63]	IJCAI19	94.83	87.34	89.36	79.16	73.12	41.38
	CAMA (N=3) [62]	CVPR19	94.70	84.50	85.80	72.90		
	HBFP-Net [33]	ACM MM20	95.80	89.80	89.50	80.20		
	Proposed LDS		95.84	90.37	91.56	82.50	86.54	67.21
+ Re-rank	VCFL [30] + Re-rank	ICCV19	90.91	86.67				
	DCDS [2] + Re-rank	ICCV19	95.40	93.30	88.50	86.10		
	Auto-ReID [46] + Re-rank	ICCV19	95.40	94.20				
	SFT [36] + Re-rank	ICCV19	93.50	90.60	88.30	83.30	76.10	60.80
	MVP Loss [48] + Re-rank	ICCV19	93.30	90.90	86.30	83.90		
	Proposed LDS + Re-rank		96.17	94.89	92.91	91.00	88.35	79.09

¹ The best results are in bold.² The metric 'R1' is the abbreviation of 'Rank-1'.

pared to the best competitor, DSFL [44], our method achieves performance improvement of 1.3% and 1.4% on the metric of Rank-1 and mAP, respectively. For the re-rank counterparts, we conduct performance improvement of 4.4% and 4.7% on the metric of Rank-1 and mAP compared to the best competitor, DCDS [2]. These comparisons demonstrate our scheme achieves considerable improvement compared to other state-of-the-art methods.

Performances on MSMT17. Table 1 also represents the proposed LDS achieves the state-of-the-art performance. For the metric of Rank-1, we achieve a performance improvement of 2.3% compared to the best competitor, DSFL [44].

For the metric of mAP, we achieve a performance improvement of 6.4% compared to the best competitor, ABD-Net [7]. For the re-rank version, our scheme achieves significant improvement in the metric of mAP, *i.e.*, 18.29%, compared to the best competitor, SFT [36]. These comparisons demonstrate the effectiveness of our method.

4.4. Ablation Study

Table 2 shows the ablation study. The baseline is a one-branch network and trained using classification loss and triplet loss, following the same backbone and training parameters with the proposed LDS. We also use DML [65] as the baseline of the multi-branch network. The evaluation of DML

Table 2

Performance comparisons with the baseline and DML.

# Branch	Method	Random Erasing	Random Scaling	Homologous Input	Market1501		DukeMTMC-reID		MSMT17	
					R1	mAP	R1	mAP	R1	mAP
1-Branch	Baseline				95.01	86.57	88.78	76.29	81.65	55.53
2-Branch	DML-2[65]†				95.34	87.76	89.59	77.91	84.30	60.30
	LDS-2 ⁽¹⁾			✓	95.55	88.28	90.44	79.21	84.92	61.62
	LDS-2 ⁽²⁾		✓		95.37	87.75	89.00	77.80	83.94	59.58
	LDS-2 ⁽³⁾		✓	✓	95.61	88.19	89.99	79.31	84.73	61.35
	LDS-2 ⁽⁴⁾	✓			95.75	89.94	90.93	81.74	86.10	66.24
	LDS-2 ⁽⁵⁾	✓		✓	95.55	90.24	90.98	82.17	86.49	67.05
3-Branch	DML-3[65]†				95.58	87.95	89.18	78.01	84.39	60.55
	LDS-3 ⁽¹⁾			✓	95.16	88.25	89.95	79.55	85.16	62.46
	LDS-3 ⁽²⁾	✓	✓		95.72	89.76	90.31	80.88	85.89	65.25
	LDS-3 ⁽³⁾	✓	✓	✓	95.84	90.37	91.56	82.50	86.54	67.21

¹ '†': reimplemented by us.² The metric 'R1' is the abbreviation of 'Rank-1'.

also uses the concatenated features from different branches. In table 2, the LDS- $i^{(j)}$ denotes the j^{th} configuration of the LDS with i branches, and the DML- i denotes the DML with i branches.

In Table 2, the proposed *homologous input* ensures the different branches have the same source image. The LDS with two branches has one master branch and one servant branch, and the LDS with three branches has one master branch and two servant branches. These servant branches are utilized to deal with occluded and scale variation scenes.

Effectiveness of Proposed Scheme. We make three comparisons to demonstrate the effectiveness of the proposed scheme. These comparisons are between the baseline and the LDS-2⁽⁵⁾, between the DML-2 and the LDS-2⁽⁵⁾, between the DML-3 and the LDS-3⁽³⁾. Compared to the baseline, the LDS-2⁽⁵⁾ configuration achieves noticeable performance improvement, *i.e.*, an average Rank-1 improvement of 2.5% and an average mAP improvement of 7% for the three benchmarks. Compared to the DML-2 [65], the LDS-2⁽⁵⁾ has an average Rank-1 score of 1.2% and an average mAP score of 4.4% superiority. Compared to DML-3, the LDS-3⁽³⁾ achieves an average Rank-1 score of 1.5% and an average mAP score of 4.5% performance improvement. These comparisons demonstrate the effectiveness of the proposed scheme over the one-branch and multi-branch networks.

Effectiveness of Proposed Random Scaling. We make another three comparisons to demonstrate the effectiveness of the proposed *random scaling*. These comparisons are between the baseline and the LDS-2⁽³⁾, between the DML-2 and the LDS-2⁽³⁾, between the LDS-2⁽⁵⁾ and the LDS-3⁽³⁾. Compared to the baseline, the LDS-2⁽³⁾ configuration achieves an average Rank-1 improvement of 1.6% and average mAP improvement of 3.4% for the three benchmarks. Compared to the DML-2, the LDS-2⁽³⁾ improves an average Rank-1 score of 0.3% and an average mAP score of 0.9%. Compared to the LDS-2⁽⁵⁾, LDS-3⁽³⁾ achieves an average

Rank-1 score of 0.3% and an average mAP score of 0.2% performance improvement. These comparisons demonstrate the proposed *random scaling* can effectively improve the performance. We also make another two comparisons, which are between the LDS-2⁽²⁾ and the DML-2, between the LDS-2⁽³⁾ and the DML-2. The LDS-2⁽²⁾ achieves inferior performance compared to the DML-2. Although introducing more scale variations, *random scaling* introduces an additional misalignment problem when zoom value is between 0.9 and 1.0. In this situation, the misalignment problem gets severer, which results in a worse performance. After applying the *homologous input*, the LDS-2⁽³⁾ achieves better performance than the DML-2.

Effectiveness of Proposed Homologous Input. We make two comparisons to demonstrate the effectiveness of the proposed *homologous input*. These comparisons are between the DML-2 and the LDS-2⁽¹⁾, between the DML-3 and the LDS-3⁽¹⁾. Compared to the DML-2, the LDS-2⁽¹⁾ employs the *homologous input* and achieves an average Rank-1 score of 0.5% and an average mAP score of 1.0% performance improvement on the three benchmarks. The comparisons between the DML-3 and LDS-3⁽¹⁾ also have similar performance. We explain these performance improvements below. In the DML, the different optimization processes of each branch are mainly derived from the random initialization of the non-local layer and the *random data augmentation*. The proposed *homologous input* makes each branch have the same source images and avoids the misalignment problem introduced by the *random data augmentation*. The random initialization of the non-local layer ensures the branches have a different learning process. Therefore, employing the *homologous input* can help to achieve a noticeable performance improvement.

Random Erasing vs. Proposed Random Scaling. Applying different servant branches brings different performance improvements. For the two-branch version in the table 2, applying servant branch for occlusion, *i.e.*, LDS-2⁽⁵⁾ achieves

Table 3

Performance comparisons on the P-DukeMTMC-reID benchmark under a supervised setting.

Method	Venue	Rank-1	Rank-5	Rank-10	mAP
PCB [50]	ECCV2018	79.4	87.1	90.0	63.9
PVPM [14]	CVPR2020	85.1	91.3	93.3	69.9
Baseline		88.2	93.1	94.3	76.4
DML-2		90.5	94.1	95.1	78.8
LDS-2 ⁽⁵⁾		91.9	95.2	96.3	82.9

Table 4

Performance comparisons on the Occluded-DukeMTMC benchmark under a supervised setting.

Method	Venue	Rank-1	Rank-5	Rank-10	mAP
PGFA [40]	ICCV2019	51.4	68.6	74.9	37.3
HOReID [55]	CVPR2020	55.1			43.8
Baseline		62.6	75.1	80.6	50.2
DML-2		63.6	76.4	80.5	51.9
LDS-2 ⁽⁵⁾		64.3	77.1	82.6	55.7

better performance than scale variation, *i.e.*, LDS-2⁽³⁾. This phenomenon may be caused by the fact that the occlusion scenes are more common than the scale variation scenes in the three benchmarks. Or the *random erasing* can increase the diversity of input images more effectively than the proposed *random scaling*.

Effectiveness in Occluded Scenes. To verify effectiveness of the proposed scheme in the occluded scenes, we train LDS-2⁽⁵⁾ on the P-DukeMTMC-reID and the Occluded-DukeMTMC benchmarks. These two large-scale benchmarks include training sets for model learning. The results are illustrated in table 3 and table 4. The baseline method and the DML with two branches are also trained on these two benchmarks. The performances on these two large-scale benchmarks demonstrate the effectiveness of the proposed scheme in the occluded scenes. In table 3, the proposed LDS 2⁽⁵⁾ achieves state-of-the-art performance on P-DukeMTMC-reID benchmark under a supervised setting. In table 3, PCB [50] is a stripe/part-related method, and PVPM [14] utilized the pose-guided attention to mine the part visibility for the occluded ReID task. Based on the strong baseline method, the proposed LDS 2⁽⁵⁾ improves the performance further. In table 4, the proposed LDS 2⁽⁵⁾ achieves state-of-the-art performance on Occluded-DukeMTMC benchmark under a supervised setting. In table 4, PGFA [40] exploited the pose landmarks to disentangle the visible region from the occlusion noise. HOReID [55] utilized the key-point information to obtain a local-feature graph to learn the high-order relation and topology knowledge. Compared with them, our method employs a simple idea and also achieves a better performance.

Effectiveness of Mutual Learning. Although multiple branches in our scheme employ mutual learning to transfer knowledge to each branch, we also propose master-servant learning (MS Learning). In MS learning, the knowledge communication is only between the master branch and ser-

Table 5

Performance comparisons between the proposed LDS using mutual learning and master-servant learning on different benchmarks.

	Market1501		DukeMTMC-reID		MSMT17	
	R1	mAP	R1	mAP	R1	mAP
DML-3 [65]	95.58	87.95	89.18	78.01	84.39	60.55
w/ MS Learning	95.64	89.89	91.34	81.97	86.27	66.25
w/ Mutual Learning	95.84	90.37	91.56	82.50	86.54	67.21

vant branch. And there is no knowledge communication between the different servant branches. In table 5, the performance comparisons between the proposed LDS using mutual learning and MS learning are listed. The LDS with MS learning achieves better performance than the DML-3. However, the LDS with MS learning is inferior to the LDS with mutual learning. We explain these performances below. For master-servant learning, the knowledge from different servant branches is indirectly communicated through the intermediate master branch. For mutual learning, the servant branches have direct communication, and the experiment results indicate this direct communication is more effective. Therefore, we apply mutual learning in our scheme to promote each branch.



Figure 5: Retrieval image examples. The query image is from the MSMT17 benchmark. The query and retrievals contain obvious occlusion and scale variation. The correct and incorrect ones are in a green and red box, respectively.

Qualitative Analysis. We show the visual comparisons of retrieval images in Fig. 5. We select a query image with occlusion from the MSMT17 benchmark and compare the retrievals of different methods, including baseline, DML, and the proposed LDS. In Fig. 5, the baseline method neglects many obvious correct results. The DML method identifies those easy image samples. However, it fails in the scale variation scenes, *e.g.*, the last result in 5(d). These results demonstrate the effectiveness of our scheme in occluded

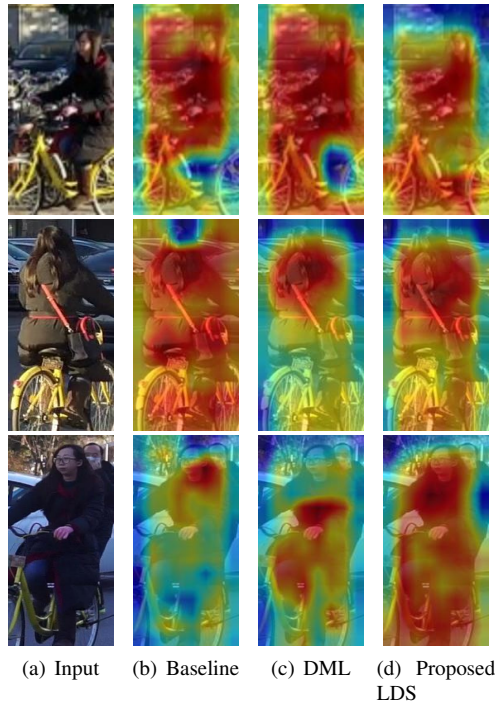


Figure 6: Activation maps of images in Fig. 5. In each row, the first one is the original input image. The second, third, and fourth ones are the grad-CAM++ [3] results from the models of baseline method, DML, and the proposed LDS, respectively.

scenes and scale variation scenes. To further analyze the learning ability of different models, we show the activation maps of different methods for the same input image, as illustrated in Fig. 6. Since the DML and the proposed LDS have three branches, we only show the activation map of the first branch. For each example, the activation map of the proposed LDS shows better responses than the others. These comparisons explain why the proposed LDS performs better and demonstrate its effectiveness.

5. Conclusion

In this paper, we propose to learn to disentangle scenes for the ReID task. This scheme employs a divide-and-conquer strategy for the ReID task. Concretely, we use two self-supervision operations to generate new image with the characteristics of occluded and scale variation scenes. Then we utilize two servant branches to deal with them. In this way, the burden of each branch is relieved. We also use a master branch to handle the general scenes. Mutual learning is employed to promote each branch. Through collaborative learning, the servant branch learns the missing information through guidance from the master branch. Moreover, the knowledge from the servant branches makes the master branch more robust. Extensive experimental results show that our method outperforms the existing one-branch and multi-branch networks and achieves state-of-the-art performances on three ReID benchmarks and two large-scale oc-

cluded ReID benchmarks. Additionally, the ablation study also validates our scheme can significantly improve performance in various scenes.

References

- [1] Ahn, S., Hu, S.X., Damianou, A., Lawrence, N.D., Dai, Z., 2019. Variational information distillation for knowledge transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9163–9171.
- [2] Alemu, L.T., Pelillo, M., Shah, M., 2019. Deep constrained dominant sets for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 9855–9864.
- [3] Chattopadhyay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N., 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. pp. 839–847.
- [4] Chen, B., Deng, W., Hu, J., 2019a. Mixed high-order attention network for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 371–381.
- [5] Chen, G., Lin, C., Ren, L., Lu, J., Zhou, J., 2019b. Self-critical attention learning for person re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9637–9646.
- [6] Chen, J., Qin, J., Yan, Y., Huang, L., Liu, L., Zhu, F., Shao, L., 2020a. Deep local binary coding for person re-identification by delving into the details, in: Proceedings of the 28th ACM International Conference on Multimedia, pp. 3034–3043.
- [7] Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z., Wang, Z., 2019c. Abd-net: Attentive but diverse person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 8351–8361.
- [8] Chen, X., Fu, C., Zhao, Y., Zheng, F., Song, J., Ji, R., Yang, Y., 2020b. Saliency-guided cascaded suppression network for person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3300–3310.
- [9] Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N., 2016. Person re-identification by multi-channel parts-based cnn with improved triplet loss function, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1335–1344.
- [10] Dai, Z., Chen, M., Gu, X., Zhu, S., Tan, P., 2019. Batch dropblock network for person re-identification and beyond, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 3691–3701.
- [11] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Ieee. pp. 248–255.
- [12] Fang, P., Zhou, J., Roy, S.K., Petersson, L., Harandi, M., 2019. Bi-linear attention networks for person retrieval, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 8030–8039.
- [13] Gao, L., Zhang, H., Gao, Z., Guan, W., Cheng, Z., Wang, M., 2020a. Texture semantically aligned with visibility-aware for partial person re-identification, in: Proceedings of the 28th ACM International Conference on Multimedia, pp. 3771–3779.
- [14] Gao, S., Wang, J., Lu, H., Liu, Z., 2020b. Pose-guided visible part matching for occluded person reid, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11744–11752.
- [15] Gao, Z., Gao, L.S., Zhang, H., Cheng, Z., Hong, R., 2019. Deep spatial pyramid features collaborative reconstruction for partial person reid, in: Proceedings of the 27th ACM International Conference on Multimedia, pp. 1879–1887.
- [16] Ge, Y., Chen, D., Li, H., 2020. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. arXiv preprint arXiv:2001.01526.
- [17] Gray, D., Brennan, S., Tao, H., 2007. Evaluating appearance models for recognition, reacquisition, and tracking, in: Proc. IEEE In-

- ternational Workshop on Performance Evaluation for Tracking and Surveillance (PETS), Citeseer. pp. 1–7.
- [18] Guo, J., Yuan, Y., Huang, L., Zhang, C., Yao, J.G., Han, K., 2019. Beyond human parts: Dual part-aligned representations for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 3642–3651.
 - [19] Hao, Y., Wang, N., Gao, X., Li, J., Wang, X., 2019. Dual-alignment feature embedding for cross-modality person re-identification, in: Proceedings of the 27th ACM International Conference on Multimedia, pp. 57–65.
 - [20] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.
 - [21] He, L., Liang, J., Li, H., Sun, Z., 2018. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7073–7082.
 - [22] He, L., Liao, X., Liu, W., Liu, X., Cheng, P., Mei, T., 2020. Fastreid: A pytorch toolbox for general instance re-identification. arXiv preprint arXiv:2006.02631 6, 8.
 - [23] Hermans, A., Beyer, L., Leibe, B., 2017. In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 .
 - [24] Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 .
 - [25] Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X., 2019. Interaction-and-aggregation network for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9317–9326.
 - [26] Huang, H., Yang, W., Chen, X., Zhao, X., Huang, K., Lin, J., Huang, G., Du, D., 2018. Eanet: Enhancing alignment for cross-domain person re-identification. arXiv preprint arXiv:1812.11369 .
 - [27] Jin, X., Lan, C., Zeng, W., Chen, Z., 2020. Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification. arXiv preprint arXiv:2001.05197 .
 - [28] Jin, X., Lan, C., Zeng, W., Wei, G., Chen, Z., 2019. Semantics-aligned representation learning for person re-identification. arXiv preprint arXiv:1905.13143 .
 - [29] Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .
 - [30] Liu, F., Zhang, L., 2019. View confusion feature learning for person re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6639–6648.
 - [31] Liu, W., Zhang, Y.M., Li, X., Yu, Z., Dai, B., Zhao, T., Song, L., 2017a. Deep hyperspherical learning. arXiv preprint arXiv:1711.03189 .
 - [32] Liu, Y., Li, H., Wang, X., 2017b. Rethinking feature discrimination and polymerization for large-scale recognition. arXiv preprint arXiv:1710.00870 .
 - [33] Liu, Z., Zhang, L., Yang, Y., 2020. Hierarchical bi-directional feature perception network for person re-identification, in: Proceedings of the 28th ACM International Conference on Multimedia, pp. 4289–4298.
 - [34] Loshchilov, I., Hutter, F., 2016. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 .
 - [35] Lu, Y., Wu, Y., Liu, B., Zhang, T., Li, B., Chu, Q., Yu, N., 2020. Cross-modality person re-identification with shared-specific feature transfer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13379–13389.
 - [36] Luo, C., Chen, Y., Wang, N., Zhang, Z., 2019a. Spectral feature transformation for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 4976–4985.
 - [37] Luo, H., Jiang, W., Zhang, X., Fan, X., Qian, J., Zhang, C., 2019b. Alignedreid++: Dynamically matching local information for person re-identification. Pattern Recognition 94, 53–61.
 - [38] Lv, J., Li, Z., Nai, K., Chen, Y., Yuan, J., 2020a. Person re-identification with expanded neighborhoods distance re-ranking. Image and Vision Computing 95, 103875.
 - [39] Lv, Y., Gu, Y., Xingqiao, L., 2020b. The dilemma of trihard loss and an element-weighted trihard loss for person re-identification. Advances in Neural Information Processing Systems 33.
 - [40] Miao, J., Wu, Y., Liu, P., Ding, Y., Yang, Y., 2019. Pose-guided feature alignment for occluded person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 542–551.
 - [41] Munjal, B., Galasso, F., Amin, S., 2019. Knowledge distillation for end-to-end person search. arXiv preprint arXiv:1909.01058 .
 - [42] Pan, X., Luo, P., Shi, J., Tang, X., 2018. Two at once: Enhancing learning and generalization capacities via ibn-net, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 464–479.
 - [43] Park, H., Ham, B., 2019. Relation network for person re-identification. arXiv preprint arXiv:1911.09318 .
 - [44] Peng, P., Tian, Y., Huang, Y., Wang, X., An, H., 2020. Discriminative spatial feature learning for person re-identification, in: Proceedings of the 28th ACM International Conference on Multimedia, pp. 274–283.
 - [45] Porrello, A., Bergamini, L., Calderara, S., 2020. Robust re-identification by multiple views knowledge distillation. arXiv preprint arXiv:2007.04174 .
 - [46] Qian, R., Dong, X., Wu, Y., Zhu, L., Yang, Y., 2019. Auto-reid: Searching for a part-aware convnet for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 3750–3759.
 - [47] Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C., 2016. Performance measures and a data set for multi-target, multi-camera tracking, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer. pp. 17–35.
 - [48] Sun, H., Chen, Z., Yan, S., Xu, L., 2019a. Mvp matching: A maximum-value perfect matching for mining hard samples, with application to person re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6737–6747.
 - [49] Sun, Y., Xu, Q., Li, Y., Zhang, C., Li, Y., Wang, S., Sun, J., 2019b. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 393–402.
 - [50] Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S., 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 480–496.
 - [51] Tian, Y., Krishnan, D., Isola, P., 2019. Contrastive representation distillation. arXiv preprint arXiv:1910.10699 .
 - [52] Wang, F., Cheng, J., Liu, W., Liu, H., 2018a. Additive margin softmax for face verification. IEEE Signal Processing Letters 25, 926–930.
 - [53] Wang, F., Xiang, X., Cheng, J., Yuille, A.L., 2017. Normface: L2 hypersphere embedding for face verification, in: Proceedings of the 27th ACM International Conference on Multimedia, pp. 1041–1049.
 - [54] Wang, G., Gong, S., Cheng, J., Hou, Z., 2020a. Faster person re-identification. arXiv preprint arXiv:2008.06826 .
 - [55] Wang, G., Yang, S., Liu, H., Wang, Z., Yang, Y., Wang, S., Yu, G., Zhou, E., Sun, J., 2020b. High-order information matters: Learning relation and topology for occluded person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6449–6458.
 - [56] Wang, X., Girshick, R., Gupta, A., He, K., 2018b. Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803.
 - [57] Wei, L., Zhang, S., Gao, W., Tian, Q., 2018. Person transfer gan to bridge domain gap for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 79–88.
 - [58] Xia, B.N., Gong, Y., Zhang, Y., Poellabauer, C., 2019. Second-order non-local attention networks for person re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3760–3769.
 - [59] Xu, B., He, L., Liao, X., Liu, W., Sun, Z., Mei, T., 2020. Black re-

- id: A head-shoulder descriptor for the challenging problem of person re-identification. arXiv preprint arXiv:2008.08528 .
- [60] Yang, F., Li, K., Zhong, Z., Luo, Z., Sun, X., Cheng, H., Guo, X., Huang, F., Ji, R., Li, S., 2020a. Asymmetric co-teaching for unsupervised cross-domain person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 12597–12604.
 - [61] Yang, T., Zhu, S., Chen, C., Yan, S., Zhang, M., Willis, A., 2020b. Mutualnet: Adaptive convnet via mutual learning from network width and resolution. arXiv preprint arXiv:1909.12978 .
 - [62] Yang, W., Huang, H., Zhang, Z., Chen, X., Huang, K., Zhang, S., 2019. Towards rich feature discovery with class activation maps augmentation for person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1389–1398.
 - [63] Yu, Z., Jin, Z., Wei, L., Guo, J., Huang, J., Cai, D., He, X., Hua, X.S., 2019. Progressive transfer learning for person re-identification. arXiv preprint arXiv:1908.02492 .
 - [64] Zhang, L., Xiang, T., Gong, S., 2016. Learning a discriminative null space for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1239–1248.
 - [65] Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H., 2018. Deep mutual learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4320–4328.
 - [66] Zhang, Z., Jiang, S., Huang, C., Li, Y., Da Xu, R.Y., 2020a. Rgb-ir cross-modality person reid based on teacher-student gan model. arXiv preprint arXiv:2007.07452 .
 - [67] Zhang, Z., Lan, C., Zeng, W., Chen, Z., 2019. Densely semantically aligned person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 667–676.
 - [68] Zhang, Z., Lan, C., Zeng, W., Jin, X., Chen, Z., 2020b. Relation-aware global attention for person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3186–3195.
 - [69] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q., 2015. Scalable person re-identification: A benchmark, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1116–1124.
 - [70] Zheng, L., Yang, Y., Hauptmann, A.G., 2016. Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 .
 - [71] Zheng, Z., Zheng, L., Yang, Y., 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 3754–3762.
 - [72] Zhong, Z., Zheng, L., Cao, D., Li, S., 2017. Re-ranking person re-identification with k-reciprocal encoding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1318–1327.
 - [73] Zhou, B., Cui, Q., Wei, X.S., Chen, Z.M., 2020a. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9719–9728.
 - [74] Zhou, J., Roy, S.K., Fang, P., Harandi, M., Petersson, L., 2020b. Cross-correlated attention networks for person re-identification. Image and Vision Computing 100, 103931.
 - [75] Zhou, J., Su, B., Wu, Y., 2020c. Online joint multi-metric adaptation from frequent sharing-subset mining for person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2909–2918.
 - [76] Zhou, K., Yang, Y., Cavallaro, A., Xiang, T., 2019a. Omni-scale feature learning for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 3702–3712.
 - [77] Zhou, S., Wang, F., Huang, Z., Wang, J., 2019b. Discriminative feature learning with consistent attention regularization for person re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8040–8049.
 - [78] Zhu, K., Guo, H., Liu, Z., Tang, M., Wang, J., 2020. Identity-guided human semantic parsing for person re-identification. arXiv preprint arXiv:2007.13467 .
 - [79] Zhuo, J., Chen, Z., Lai, J., Wang, G., 2018. Occluded person re-identification, in: 2018 IEEE International Conference on Multimedia and Expo (ICME), IEEE. pp. 1–6.
 - [80] Zhuo, J., Lai, J., Chen, P., 2019. A novel teacher-student learning framework for occluded person re-identification. arXiv preprint arXiv:1907.03253 .