Pose-Guided Feature Learning with Knowledge Distillation for Occluded Person Re-Identification

Kecheng Zheng[1], Cuiling Lan[2*], Wenjun Zeng[2], Jiawei Liu[1], Zhizheng Zhang[2], Zheng-Jun Zha[1*]
[1]University of Science and Technology of China
[2]Microsoft Research Asia
{zkcys001,jwliu6}@mail.ustc.edu.cn,{culan,wezeng,zhizzhang}@microsoft.com,zhazj@ustc.edu.cn

ABSTRACT

Occluded person re-identification (ReID) aims to match person images with occlusion. It is fundamentally challenging because of the serious occlusion which aggravates the misalignment problem between images. At the cost of incorporating a pose estimator, many works introduce pose information to alleviate the misalignment in both training and testing. To achieve high accuracy while preserving low inference complexity, we propose a network named Pose-Guided Feature Learning with Knowledge Distillation (PGFL-KD), where the pose information is exploited to regularize the learning of semantics aligned features but is discarded in testing. PGFL-KD consists of a main branch (MB), and two pose-guided branches, *i.e.*, a foreground-enhanced branch (FEB), and a body part semantics aligned branch (SAB). The FEB intends to emphasise the features of visible body parts while excluding the interference of obstructions and background (*i.e.*, foreground feature alignment). The SAB encourages different channel groups to focus on different body parts to have body part semantics aligned representation. To get rid of the dependency on pose information when testing, we regularize the MB to learn the merits of the FEB and SAB through knowledge distillation and interaction-based training. Extensive experiments on occluded, partial, and holistic ReID tasks show the effectiveness of our proposed network.

CCS CONCEPTS

• **Computing methodologies** → **Object identification**.

KEYWORDS

Occluded Person Re-Identification, Human Pose, Knowledge Distillation, Feature Alignment

ACM Reference Format:
Kecheng Zheng[1], Cuiling Lan[2*], Wenjun Zeng[2], Jiawei Liu[1], Zhizheng Zhang[2], Zheng-Jun Zha[1*]. 2021. Pose-Guided Feature Learning with Knowledge Distillation for Occluded Person Re-Identification. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475610

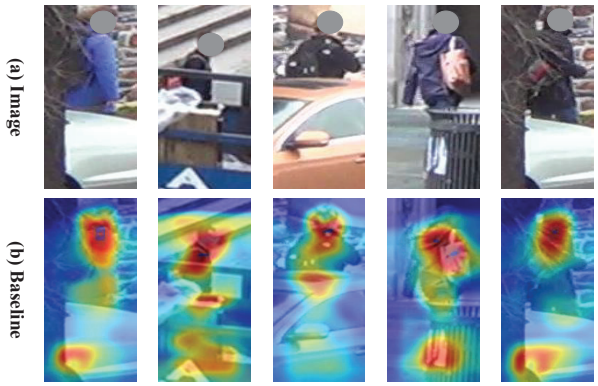

Figure 1: Examples of (a) occluded/partial person images and (b) the feature responses of Baseline. In (b), for the regions with objects occluding persons (i.e., obstructions), the networks usually mistakenly generate high responses by regarding them as discriminative person regions.

1 INTRODUCTION

Person re-identification (ReID) [12, 14, 25–27, 49, 50, 52, 53], aims to match images of a person across cameras, which has many applications such as person tracking in a retail store, finding lost child, *etc.* In recent years, many methods have been proposed for person ReID [11, 17, 20, 22, 41, 43, 44, 54]. However, most of them focus on holistic person images and only very few works investigate the more challenging occluded person ReID [2, 7, 10, 29, 30, 40, 42, 46, 60], even though the occluded person images are very common in practical scenarios. As shown in Figure 1, a person is usually occluded by some objects (*e.g.* tree, car, board, bucket) or walks out of the camera field, leading to occluded or partial person images.

Occluded/partial person ReID is challenging, where there are both occluded/partial person images and holistic person images for matching [1]. First, the spatial misalignment between an occluded/partial person image and a holistic person image or between two occluded/partial person image is in general more severe than that between two holistic person images. Second, as examples shown in Figure 1 (b), for the regions with objects occluding persons (*i.e.*, obstructions), the networks usually mistakenly generate

This work was done when Kecheng was an intern at MSRA.
*Corresponding authors: Cuiling Lan, Zheng-Jun Zha.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475610

[1]Note that, actually, the partial person image can be also considered as the occluded person image in a broad sense where the "occluded" region is not presented in the image.

high responses by regarding them as discriminative person regions, resulting in interference to the person feature representation.

Recently, some occluded/partial person ReID methods are proposed [2, 7, 10, 29, 30, 40, 42, 47, 60]. Many works alleviate the misalignment by learning both global and local body part features [2, 30, 42] for matching. Matching based on local body part (*e.g.*, head, arm, leg, *etc.*) features facilitates the matching between two regions of the same semantics, alleviating the matching difficulty from spatial misalignment. Besides, such decoupling of body parts could confine the interference caused by the missing of some body parts into the local features rather than distributed to the global scope feature. On the other hand, a local body part feature may not be capable of capturing some attributes which require a more global observation (*e.g.*, a person is wearing the clothes of the same color for upper body and lower body). **Thus, both global information and local information are vital especially for occluded person ReID. However, in general, to extract local body part features, an external pose estimator is utilized in both training and testing. This increases the complexity of the model (*i.e.*, model size, computational cost) in testing/inference and is not friendly in deployment.**

In this work, we aim to preserve the merits of the global features and local body part features for occluded person re-identification, while eliminating the requirement of a pose estimator in testing for low complexity. To this end, we propose a Pose-Guided Feature Learning with Knowledge Distillation (PGFL-KD) network, where the pose information is exploited to regularize the learning of global features and the pose estimator is discarded in testing. PGFL-KD consists of a main branch (MB), and two pose-guided branches: a foreground-enhanced branch (FEB), and a body part semantics aligned branch (SAB). First, we explicitly alleviate the interference from the obstructions (see Figure 1 (b)) by learning foreground-enhanced feature in the FEB, where we define the foreground as the regions around detected visible body joints based on pose. Second, based on pose, we enable different channel groups to represent features of the different body parts to have semantics aligned representation in the SAB. To get rid of the dependency on pose information when testing, the MB is "taught" to ignore the interference from obstructions and background through knowledge distillation, and to learn semantics aligned representations through our interaction-based training, where the latter is promoted by our multi-part contrastive loss and interaction-based training.

The main contributions of this paper are summarized as follows:

- We propose a Pose-Guided Feature Learning with Knowledge Distillation (PGFL-KD) network for effective occluded person re-identification. Through pose-guided interaction learning (*i.e.*, knowledge distillation and interaction-based training), we enable the discarding of dependency on pose estimator while preserving high performance in testing.
- We introduce two pose-guided branches in the training in order to possess two merits for teaching the MB: 1) exclusion of the interference from the obstructions and background (by the FEB); 2) semantics aligned feature representation (by the SAB).

To the best of our knowledge, this is the first work that distills the robust feature representations based on pose in training but does not need pose in testing for occluded person ReID. Extensive experiments on occluded, partial, and holistic ReID tasks show the effectiveness of our proposed network and validate the superiority of PGFL-KD over various state-of-the-art methods.

2 RELATED WORKS

Occluded/Partial Person Re-identification. Occluded person ReID [60] aims to match person images of both occluded/partial person and holistic person. Zheng *et al.* [56] propose a global-to-local matching model to capture the spatial layout information. He *et al.* [7] reconstruct the feature map of a partial query from the holistic pedestrian, and further improve it with a foreground-background mask to reduce the influence of background clutter in [10]. Iodice *et al.* [15] align partial views by using human pose information and hallucinate the missing parts with a Cycle-Consistent Adversarial Networks [58]. Sun *et al.* propose a Visibility-aware Part Model (VPM) in [40], which learns to perceive the visibility of regions by self-supervised learning. Zhuo *et al.* [60] propose occluded/non-occluded binary classification (OBC) loss to regularize the feature learning. Luo *et al.* [29] propose a spatial transform module to transform the holistic image to align with the partial ones, and further calculate the distance of the aligned pairs. Fan *et al.* [1] propose a spatial-channel parallelism network (SCPNet) that encodes spatial body part features into specific channels and fuses the holistic and part features to obtain discriminative features. However, the spatial parts are obtained by dividing the feature map into several spatial horizontal stripes which cannot assure the alignment of body part semantics. Recently, many works introduce a pose estimator [39, 48] to obtain semantics aligned local body part features, and global feature for matching [2, 30, 42]. Miao *et al.* [30] propose a pose guided feature alignment method, which extracts the local body part features based on pose and alleviates the influence of occluded body regions in matching. HOReID [42] adopts high-order relation and human-topology information for feature learning and alignment. PVPM [2] utilizes the characteristic of part correspondence to estimate whether a part suffers from the occlusion or not. However, these methods extract local body part features based on pose information and a pose estimator is needed in testing, which increases the complexity of inference model.

In this paper, we aim to exploit the local body part semantics aligned feature representations for high performance while discarding the dependency on a pose estimator in testing for low complexity.

Knowledge Distillation. Knowledge distillation is one of the most popular techniques in model compression and acceleration [4]. It in general transfers knowledge from one model (*i.e.*, a teacher) to another (*i.e.*, a student), usually from a larger model to a smaller one. In this work, we aim to transfer the knowledge from the pose-guided branches to the main branch during the training, which enables the discarding of pose-based branches while maintaining good performance.

3 PROPOSED METHOD

We propose a network named Pose-Guided Feature Learning with Knowledge Distillation (PGFL-KD) for occluded person ReID. Figure 2 shows the flowchat for training. PGFL-KD consists of three branches: a main branch (MB), a foreground-enhanced branch (FEB), and a body part semantics aligned branch (SAB). Guided by pose

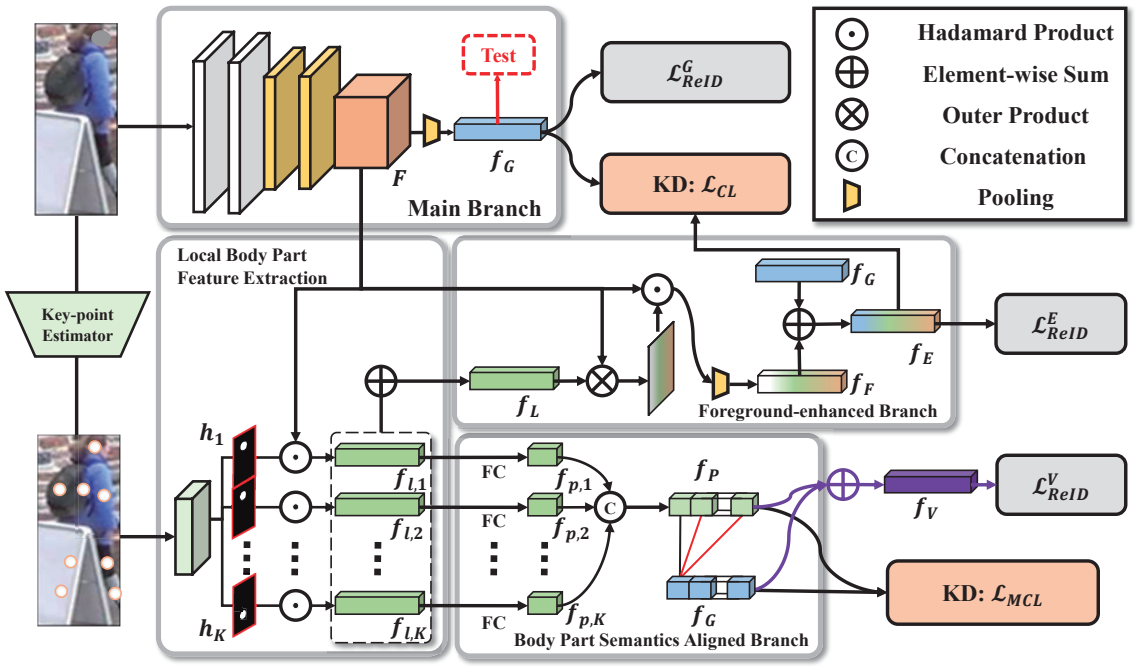

Figure 2: Illustration of our proposed network Pose-Guided Feature Learning with Knowledge Distillation (PGFL-KD), where the pose information is exploited to regularize the learning of semantics aligned features but is discarded in testing. PGFL-KD consists of a main branch (MB) and two pose-guided branches: a foreground-enhanced branch (FEB), and a body part semantics aligned branch (SAB). In testing, only the MB is needed. The FEB aims to alleviate the interference of obstructions and background by learning foreground-enhanced feature. The SAB aims to learn body part semantics aligned feature representations. We distill the knowledge from the two branches to the MB by the knowledge distillation losses (*i.e.*, consistent loss w.r.t. the FEB, and multi-part contrastive loss and interaction-based training (marked by purple) w.r.t. the SAB.

information, the FEB learns foreground-enhanced feature which allevates the interference from obstructions and background (see Sec. 3.2) while the SAB learns body part semantics aligned feature (see Sec. 3.3). We promote the global feature in the MB to possess the merits of the features in the other two branches by distilling knowledge from them. Particularly, we encourage the global feature f_G to approach the foreground-enhanced feature f_E by adding consistent loss. Moreover, we encourage the global feature $\mathbf{f}_G$ to be body part semantically aligned as $\mathbf{f}_P$ by using multi-part contrastive loss and enabling channel wise fusion. In this way, we are capable of exploiting only the MB in the testing with satisfied performance, where the pose related two branches are discarded.

3.1 Feature Extraction

Similar to others works, we exploit both global features and local body part features in training. In contrast, we leverage the local body part features to regularize the global feature learning and only use the global feature in inference. We review how to obtain them.

Global Feature Extraction. As illustrated in Figure 2, for the main branch, we use a backbone network (*e.g.*, ResNet-50) to extract a feature map $F \in \mathbb{R}^{h\times w\times c}$, where h, w, c denote the height, width, and the number of channels, respectively. Then we adopt a global average pooling operation $g(\cdot)$ on the feature map F to output a global feature $f_G = g(F) \in \mathbb{R}^c$, where $c = 2048$.

Local Body Part Feature Extraction. For local body part feature extraction, we obtain local body part features with the guidance of the estimated human pose. Based on the off-the-shelf human pose (key-points) estimator (HR-Net [39]), given an input image, we obtain the heat map, with the responses identifying the estimated positions of each key point (in total 17 key points, with one channel denoting background), respectively. We merge the key-points based on semantics to have a merged heatmap $H \in \mathbb{R}^{h\times w\times K}$ of $K = 8$ channels corresponding to K key-point groups: including head, left lower arm, right lower arm, left knee, right knee, left ankle, right ankle, and torso. We denote the k^{th} channel of H as $H_k = H(:, :, k)$. To suppress noise and outliers, H_k is obtained by spatially normalizing original key-point heatmap with a softmax function. Note that for an occluded key point, the pose estimator in general output low responses in the heatmap.

With the guidance of key-points regions, we get K groups of semantic local features $\{f_{l,k}\}|_{k=1}^K$ by spatially pooling the feature with each key-point heatmap as attention, respectively. We obtain the features as:

$$f_{l,k} = g(F \odot (H_k \otimes \boldsymbol{e}_K)), \quad k = 1, \cdots, K, \tag{1}$$

where $\otimes$ denotes outer product, $\odot$ denotes Hadamard product, $g(\cdot)$ denotes global average pooling, $\boldsymbol{e}_K \in \mathbb{R}^K$ denotes a vector of all ones, $f_{l,k} \in \mathbb{R}^c$.

Based on local body part features, we explicitly alleviate the interference from the obstructions and background by learning foreground-enhanced feature in the FEB, where we define the foreground as the regions around detected visible body joints based on pose. Meanwhile, we also use the local body part features to enable different channel groups to represent features of the different body parts to have semantics aligned representation in the SAB. We introduce the FEB and SAB in details in the following subsections.

3.2 Foreground-enhanced Branch (FEB)

The Foreground-enhanced Branch (FEB) intends to emphasise the features of visible body parts while excluding the interference of occluding objects and background. Specifically, we use the sum of local body part features as a query to find more salient foreground regions in the feature map to obtain enhanced foreground feature. We distill the knowledge from the FEB to the MB.

Pose-Guided Feature Enhancement. With the local body part features $\boldsymbol{f}_{l,k}$ $(k = 1, \cdots, K)$ representing the informative foreground human parts, we intend to let the feature learning focus on more semantically meaningful regions. We propose a pose-guided foreground-enhanced module to improve the quality of the pooled feature by emphasising the features of visible body parts.

As illustrated in Figure 2, given a feature map $F \in \mathbb{R}^{h\times w\times c}$ and its pose-based pooled feature vector $\boldsymbol{f}_L \in \mathbb{R}^c$, where $\boldsymbol{f}_L = (\sum_{k=1}^{K} \boldsymbol{f}_{l,k})/K$, we first calculate the cosine similarity between the feature $\boldsymbol{f}_L$ and the feature map F at each pixel. Then we use softmax function to calculate the attention score map with the score value a position (i, j) as

$$a_{i,j} = \frac{\exp(F_{i,j} \cdot \boldsymbol{f}_L)}{\sum_{i,j} \exp(F_{i,j} \cdot \boldsymbol{f}_L)}, \quad i \in [1, h], j \in [1, w], \tag{2}$$

where $F_{i,j} \in \mathbb{R}^c$ denotes the feature vector at position (i, j) of the feature map F. After obtaining the attention score map, we use it as the weights for attentive pooling of the feature map to output the foreground feature vector as

$$\boldsymbol{f}_F = \sum_{i,j} a_{i,j} \times F_{i,j}. \tag{3}$$

This attentive pooling procedure can effectively shift the focus of the pooled feature vector to the body part regions, leading to more meaningful foreground representation. In order to enable the model to preserve more complete information, we add the foreground feature and the global feature to have the foreground-enhanced feature $\boldsymbol{f}_E = \boldsymbol{f}_F + \boldsymbol{f}_G$.

Knowledge Distillation. In order to get rid of the dependency on pose information in testing and inherent the merit of the foreground-enhanced feature, we regularize the feature learning of the MB by distilling knowledge from the FEB using consistent loss as

$$\mathcal{L}_{CL} = \|\boldsymbol{f}_G - \boldsymbol{f}_E\|_2^2, \tag{4}$$

where the feature $\boldsymbol{f}_E$ from the FEB acts as the teacher and the MB as the student (by detaching the foreground-enhanced feature $\boldsymbol{f}_E$).

3.3 Body Part Semantics Aligned Branch (SAB)

The SAB encourages different channel groups to focus on different body parts to have semantics aligned representation. In order to get rid of the dependency on pose information when testing, we regularize the MB to learn the merits of the SAB branches through multi-part contrastive loss for knowledge distillation and interaction-based training.

For the SAB, to generate body part semantics aligned feature representation $\boldsymbol{f}_P$, we reduce the number of dimension of each local body part feature $\boldsymbol{f}_{l,k}$ by K to have $\boldsymbol{f}_{p,k}$, and concatenate them as $\boldsymbol{f}_P = [\boldsymbol{f}_{p,1}, \boldsymbol{f}_{p,2}, \cdots, \boldsymbol{f}_{p,K}]$. Here $\boldsymbol{f}_{p,k} = ReLU(BN(W_k \boldsymbol{f}_{p,k}))$, where $W \in \mathbb{R}^{\frac{c}{K}\times c}$ and BN denotes batch normalization operation.

Multi-part Contrastive Loss. We use the multi-part contrastive loss to explicitly align the global feature $\boldsymbol{f}_G$ with the local body part feature $\boldsymbol{f}_P$ to have the semantics aligned feature representation.

To align with the local part feature $\boldsymbol{f}_P$, we split the global feature $\boldsymbol{f}_G$ of the MB into K groups, *i.e.*, $\boldsymbol{f}_G = [\boldsymbol{f}_{g,1}, \boldsymbol{f}_{g,2}, \cdots, \boldsymbol{f}_{g,K}]$, where $\boldsymbol{f}_{g,k} \in \mathbb{R}^{c/K}$. Particularly, for an input image, we encourage the consistency of the features between a local part feature $\boldsymbol{f}_{p,k}$ and its corresponding channel-group $\boldsymbol{f}_{g,k}$ of the global feature, and encourage the dissimilarity of the features between a local part $\boldsymbol{f}_{p,k}$ and a channel-group $\boldsymbol{f}_{g,i}$ of a different body part of the global feature, where $i \neq j$. To exploit the symmetry of a human body, we consider that feature of the left part (*e.g.*, left shoulder) of the body should also be close to the feature of the corresponding right part (*e.g.*, right shoulder). By following the design of the multi-positive contrastive loss [5], for an image, we have the multi-part contrastive loss as

$$\mathcal{L}_{MCL} = -\sum_{i=1}^{K} \log \frac{\sum_{j\in\mathcal{P}(i)} \exp(\boldsymbol{f}_{g,j} \cdot \boldsymbol{f}_{p,i})}{\sum_{j\in\mathcal{P}(i)} \exp(\boldsymbol{f}_{g,j} \cdot \boldsymbol{f}_{p,i}) + \sum_{j\in\mathcal{N}(i)} \exp(\boldsymbol{f}_{g,j} \cdot \boldsymbol{f}_{p,i})}, \tag{5}$$

where $\mathcal{N}(i)$ denotes the negative set within the global feature $\boldsymbol{f}_G$ *w.r.t.* the i^{th} part feature $\boldsymbol{f}_{p,i}$. For example, when the i^{th} part feature $\boldsymbol{f}_{p,i}$ denotes the feature of left foot, the left foot and right foot features in the global feature belongs to positive set while other part features belongs to negative set. Note that in such distillation, the local body part features act as teacher and the MB acts as student (by detaching the local body part feature $\boldsymbol{f}_P$).

Interaction-based Optimization. Besides the above multi-part contrastive loss for distilling semantics aligned representation, we enable the interaction between local part features and global feature to promote the channel-wise semantic alignment through joint learning by fusing.

Particularly, we enable the joint optimization of the two branches by fusing the global features with the local part features by element-wise addition, *i.e.*,

$$\boldsymbol{f}_V = \boldsymbol{f}_G + \boldsymbol{f}_P. \tag{6}$$

The widely-used ReID loss $\mathcal{L}_{ReID}$ (*i.e.*, the cross-entropy loss for identity classification (ID Loss), and the ranking loss of triplet loss with batch hard mining [11] (Triplet Loss)), is added on the fused feature $\boldsymbol{f}_V$, which we refer to $\mathcal{L}_{ReID}^{V}$.

The fusion followed by supervision plays the role of assisting the feature alignment, which drives different channel groups of the global feature to focus on different human body parts. We will give the analysis below.

Analysis from Perspective of Gradients: For the body part feature $\boldsymbol{f}_P$, features of different local body parts are allocated into different channel groups of $\boldsymbol{f}_P$ and are thus semantically aligned across two images. Global feature $\boldsymbol{f}_G$ is not naturally semantically aligned but it

contains more comprehensive information. We promote their interaction by element-wisely fusing them. We analyze the optimization gradients for the two branches below.

We take the triplet-loss as an example to analyze the gradients for the two branches, where the analysis w.r.t the classification loss is similar. We denote the two branch fused features of an anchor sample, a positive sample, and a negative sample as $\boldsymbol{v}_a = \boldsymbol{f}_G^a + \boldsymbol{f}_P^a$, $\boldsymbol{v}_p = \boldsymbol{f}_G^p + \boldsymbol{f}_P^p$, $\boldsymbol{v}_n = \boldsymbol{f}_G^n + \boldsymbol{f}_P^n$ respectively (that could be sampled from a mini-batch).

We define the triplet loss on the positive-pair and the negative-pair as

$$\mathcal{L}_{tri} = -\log \frac{e^{\boldsymbol{v}_a^T \cdot \boldsymbol{v}_p}}{e^{\boldsymbol{v}_a^T \cdot \boldsymbol{v}_p} + e^{\boldsymbol{v}_a^T \cdot \boldsymbol{v}_n}} = -\log\left(1 + e^{\boldsymbol{v}_a^T \cdot \boldsymbol{v}_n - \boldsymbol{v}_a^T \cdot \boldsymbol{v}_p}\right). \quad (7)$$

For the triplet loss, the gradients *w.r.t.* the two features are as

$$\frac{\partial \mathcal{L}_{tri}}{\partial \boldsymbol{f}_G^a} = \frac{\partial \mathcal{L}_{tri}}{\partial \boldsymbol{v}_a} \cdot \frac{\partial \boldsymbol{v}_a}{\partial \boldsymbol{f}_G^a} = \frac{\boldsymbol{v}_p - \boldsymbol{v}_n}{1 + e^{(\boldsymbol{f}_G^a + \boldsymbol{f}_P^a)^T \cdot \boldsymbol{v}_p - (\boldsymbol{f}_G^a + \boldsymbol{f}_P^a)^T \cdot \boldsymbol{v}_n}}, \quad (8)$$

$$\frac{\partial \mathcal{L}_{tri}}{\partial \boldsymbol{f}_P^a} = \frac{\partial \mathcal{L}_{tri}}{\partial \boldsymbol{v}_a} \cdot \frac{\partial \boldsymbol{v}_a}{\partial \boldsymbol{f}_P^a} = \frac{\boldsymbol{v}_p - \boldsymbol{v}_n}{1 + e^{(\boldsymbol{f}_G^a + \boldsymbol{f}_P^a)^T \cdot \boldsymbol{v}_p - (\boldsymbol{f}_G^a + \boldsymbol{f}_P^a)^T \cdot \boldsymbol{v}_n}}. \quad (9)$$

We can see that the gradient for each branch/feature is related with/influenced by the feature of the other branch, which denotes they are not independent but interacted. Moreover, the optimization direction (gradient) *w.r.t.* the global feature and that *w.r.t.* the part-aligned local feature are the same. When their optimization directions are the same, the two features share similar behaviors and are prone to have consistent characteristics/semantics. Specifically, the local feature is semantically aligned and thus encourages the global feature to be similarly semantically aligned.

3.4 Overall Loss Function

To drive both global branch and local branch to learn discriminative feature representations, we add the ReID loss on the global feature $\boldsymbol{f}_G$ (denoted as $\mathcal{L}_{ReID}^G$), the foreground-enhanced feature $\boldsymbol{f}_E$ (denoted as $\mathcal{L}_{ReID}^E$), and the body part semantics aligned feature (after fusion) $\boldsymbol{f}_V$ (denoted as $\mathcal{L}_{ReID}^V$). Together with the distillation losses, the overall loss is as

$$\mathcal{L} = \mathcal{L}_{ReID}^G + \mathcal{L}_{ReID}^E + \mathcal{L}_{ReID}^V + \lambda_{cl}\mathcal{L}_{CL} + \lambda_{mcl}\mathcal{L}_{MCL}, \quad (10)$$

where λ_{cl} and λ_{mcl} denote hyper-parameters for balancing the losses.

3.5 Inference/Testing

In the testing phase, we discard the pose estimator and only use the main branch (MB), where the feature $\boldsymbol{f}_G$ is used for matching. This removes the dependency on a pose estimator and makes the model simple with low computational complexity in testing.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

We evaluate our methods using four person ReID datasets, including two occluded datasets (Occluded-Duke [30], and Occluded-ReID [60]), three partial datasets (Partial-REID [7], Partial-iLIDS [7], and our generated Partial-Duke), and two holistic datasets (DukeMTMC-reID [33] and Market-1501 [55]), with details shown in Table 1.

Dataset	Train Nums (ID/Image)	Testing Nums (ID/Image)	
		Gallery	Query
Market-1501 [55]	751/12,936	750/19,732	750/3,368
DukeMTMC-reID [33]	702/16,522	1,110/17,661	702/2,228
Occluded-Duke [30]	702/15,618	1,110/17,661	519/2,210
Occluded-ReID [60]	-	200/1,000	200/1,000
Partial-REID [7]	-	60/300	60/300
Partial-iLIDS [7]	-	119/119	119/119
Partial-Duke (Ours)	702/16,522	1,110/17,661	702/2,228

Table 1: Dataset details. We evaluate our proposed method on seven public datasets, including two occluded datasets, three partial datasets and two holistic ones.

Occluded Person ReID Datasets. These datasets focus more on occluded person images, where in a cropped person image, a person is usually occluded by some other objects/obstructions. Occluded-Duke [30] is generated from DukeMTMC-reID by leaving occluded images and filtering out some noisy images. It contains 15,618 training images, 17,661 gallery images, and 2,210 occluded query images. Occluded-ReID [60] is captured by the mobile camera, consisting of 2000 images of 200 occluded persons. Each identity has five full-body person images and five occluded person images with different types of severe occlusions.

Partial Person ReID Datasets. These datasets focus more on partial person images, where only a partial of a person is presented in the image due to imperfect detection or out of camera field. In a broad sense, these are also occluded person images. Partial-REID [7] includes 600 images from 60 people, with five full-body images and five partial images per person, which is only used for the test. Partial-iLIDS [7] is based on the iLIDS [7] dataset and contains a total of 238 images from 119 people captured by multiple non-overlapping cameras in the airport, and their occluded regions are manually cropped.

The existing partial person ReID datasets are too small for reliable training and testing. For example, Partial-REID [7] includes only 600 images from 60 people and Partial-iLIDS [7] includes only 238 images from 119 people. There is a lack of large size partial person ReID dataset. To facilitate the investigation and evaluation, we generate a large partial person ReID dataset based on DukeMTMC-reID. We refer to it as Partial-Duke. The original DukeMTMC-reID dataset is not designed for the investigation/evaluation of partial person ReID due to its small number of partial person images. *We manually generate the Partial-Duke dataset.* Partial-Duke contains 50% partial images and 50% holistic images for the training/query/gallery sets. For these partial images, a half of them are the cropped upper half (prone to be the upper body) of the original images, and another half of images are the cropped upper one third of the original images (prone to be the upper body). In total, it contains 702 identities of 16,522 training images, 702 identities of 2,228 queries, and 1110 identities of 17,661 gallery images.

Holistic Person ReID Datasets. Market-1501 [55] and DukeMTMC-reID [33] are two widely-used large-scale holistic ReID datasets.

Evaluation Metrics. We use standard metrics in most person ReID literature, namely Cumulative Matching Characteristic Rank-1/5/10 (*i.e.*, R1/R5/R10) and mean average precision (mAP).

Methods	Occluded-Duke		Occluded-REID	
	Rank-1	mAP	Rank-1	mAP
DIM [45]	21.5	14.4	-	-
Part-Aligned [51]	28.8	20.2	-	-
HACNN [21]	34.4	26.0	-	-
Random Erasing [57]	40.0	30.0	-	-
PCB [41]	42.6	33.7	41.3	38.9
AFPB[60]	-	-	68.2	-
Part Bilinear [38]	36.9	-	-	-
FD-GAN [3]	40.8	-	-	-
AMC+SWM [56]	-	-	31.2	27.3
DSR [7]	40.8	30.4	72.8	62.8
SFR [9]	42.3	32	-	-
Ad-Occluded [13]	44.5	32.2	-	-
TCSDO [61]	-	-	73.7	77.9
FPR [10]	-	-	78.3	68.0
PGFA w/o pose [30]	46.0	34.4	-	-
PGFA [30]	51.4	37.3	-	-
PVPM [2]	-	-	66.8	59.5
PVPM+Aug [2]	-	-	70.4	61.2
HOReID [42]	55.1	43.8	80.3	70.2
ISP* [59]	62.8	52.3	-	-
Baseline	52.7	45.9	73.6	61.5
PGFL-KD (Ours)	**63.0**	**54.1**	**80.7**	**70.3**

Table 2: Comparison with state-of-the-arts on two occluded datasets, *i.e.* Occluded-Duke [30] and Occluded-REID [60]. * denotes that ISP [59] uses HRNet-W32 as the backbone and all other methods use ResNet50 backbone.

4.2 Implementation Details

For our PGFL-KD, We use ResNet50 pre-trained on ImageNet [34] as our backbone network. Similarly, we build our baseline scheme Baseline using ResNet50. As [28], we perform data augmentation of randomly erasing [57], cropping, and flipping. The images are resized to 384×128. Each mini-batch contains 64 images of 4 identities, where there are 16 images for each identity. Adam [19] optimizer is adopted to optimize the networks. The initial learning rate is set to 0.00035. For the identity classifiers, a BNNeck is adopted, which contains a batch normalization layer [16], and a fully connected layer followed by a softmax function. The network is jointly trained end-to-end for 120 epochs with an initialized learning rate of 3.5e-4. The learning rate is decayed by 0.1 at 30 and 70 epochs. We implement our framework with Pytorch.

The HR-Net [39] trained on the COCO dataset [23] is used to extract the human key-points. The keypoint extractor predicts 17 key-points, and we merge these key-points according to the body semantics to obtain $K = 8$ key-points. Specifically, torso consists of left/right shoulders and hips. We merge the left (or right) elbow and wrist as the left (or right) lower arm. After merging, the $K = 8$ key-points consist of head, left lower arm, right lower arm, left knee, right knee, left ankle, right ankle, and torso.

4.3 Comparison with the State-of-the-Arts

Results on Occluded Person ReID Datasets. As are shown in Table 2, we mainly compare with methods of four categories: vanilla holistic ReID methods [41, 51], holistic ReID methods with key-point information [3, 38], partial ReID methods [7, 9, 56], and occluded ReID methods [10, 13, 30].

Methods	Partial-REID		Partial-iLIDS	
	Rank-1	Rank-3	Rank-1	Rank-3
DSR [7]	50.7	70.0	58.8	67.2
SFR [9]	56.9	78.5	63.9	74.8
VPM [40]	67.7	81.9	65.5	74.8
PGFA [30]	68.0	80.0	69.1	80.9
AFPB [60]	78.5	-	-	-
FPR [10]	81.0	-	68.1	-
HOReID [42]	**85.3**	**91.0**	72.6	86.4
PGFL-KD (Ours)	85.1	90.8	**74.0**	**86.7**

Table 3: Comparison with state-of-the-art approaches on two partial datasets, *i.e.* Partial-REID [56] and Partial-iLIDS [7] datasets. Our method achieves competitive performance on the two partial datasets.

Methods	Partial-Duke			
	Rank-1	Rank-5	Rank-10	mAP
FPR [10]	69.2	83. 4	87.6	50.5
PGFA [30]	66.2	81.5	85.4	42.5
PVPM [2]	74.6	83.7	88.9	57.3
HOReID [42]	77.6	86.3	90.9	59.0
Baseline	70.1	82.2	87.7	51.2
PGFL-KD (Ours)	**81.1**	**89.5**	**92.7**	**64.2**

Table 4: Performance comparison (%) with the state-of-the-arts on our created large partial dataset, *i.e.* Partial-Duke.

The first two category approaches achieve less satisfactory results, because they do not design the networks specific to the occluded ReID. For partial/occluded ReID methods, an obvious improvement is achieved on the two datasets. Our proposed PGFL-KD achieves the best performance when compared with these state-of-the-art methods, which outperforms the second best method HOReID [42] by **10.3%** in mAP accuracy on the large dataset Occluded-Duke. Note that HOReID needs a pose estimator in testing but we do not. At the same inference complexity, our PGFL-KD outperforms the baseline scheme Baseline significantly by **8.2%** and **8.8%** in mAP on Occluded-Duke and Occluded-REID, respectively.

Results on Partial Person ReID Datasets. To further evaluate our proposed scheme, in Table 3 we report the results on two partial person ReID datasets, Partial-REID [56] and Partial-iLIDS [7]. As we can see, our proposed PGFL-KD outperforms the other methods by at least 1.4% in terms of Rank-1 accuracy on Partial-iLIDS and achieves the competitive results to HOReID[6] on Partial-REID. Our inference model is simple and does not need pose estimator but HOReID requires.

The existing partial person ReID datasets are too small for reliable training and testing. Thus we manually generate the Partial-Duke dataset, which is much larger than Partial-REID and Partial-iLIDS (see Table 1. Table 4 shows the comparison with the state-of-the-art approaches on this large Partial-Duke dataset, where the results are obtained by running their source codes. We can see that our proposed PGFL-KD achieves the best performance, which outperforms the second best method by **5.2%** in mAP accuracy.

Results on Holistic Person ReID Datasets. In considering the practical applications where both occluded and holistic person matching is needed, it is expected that a method designed for

Methods	Market-1501		DukeMTMC	
	Rank-1	mAP	Rank-1	mAP
PCB [41]	92.3	77.4	81.8	66.1
VPM [40]	93.0	80.8	83.6	72.6
BOT [28]	94.1	85.7	86.4	76.4
GCP [31]	95.2	**88.9**	87.9	78.6
SPReID [18]	92.5	81.3	-	-
MGCAM [36]	83.8	74.3	46.7	46.0
MaskReID [32]	90.0	75.3	-	-
ISP [59]	-	-	88.7	78.9
PDC [37]	84.2	63.4	-	-
Pose-transfer [24]	87.7	68.9	30.1	28.2
PSE [35]	87.7	69.0	27.3	30.2
PGFA [30]	91.2	76.8	82.6	65.5
HOReID [6]	94.2	84.9	86.9	75.6
GASM [8]	**95.3**	84.7	88.3	74.4
Baseline	94.0	85.2	86.3	76.1
PGFL-KD (Ours)	**95.3**	87.2	**89.6**	**79.5**

Table 5: Comparison with state-of-the-arts on two holistic datasets, Market-1501 and DukeMTMTc-reID.

occluded person ReID should work for holistic person ReID. We compare with the state-of-the-art approaches on holistic person ReID in Table 5. We also compare with the vanilla ReID methods [28, 40, 41], the ReID methods with human-parsing information [18, 32, 36, 59], and the holistic ReID methods with key-points information [8, 24, 30, 35, 37].

We can see that our proposed PGFL-KD achieves the competitive results on the holistic person ReID datasets. It is mentioned that our model uses only the vanilla ResNet model in testing, which does not introduce additional computational complexity and does not need a pose estimator.

4.4 Ablation Studies

In this section, we conduct ablation studies to evaluate the effectiveness of designs in the proposed PGFL-KD. PGFL-KD consists of a main branch (MB), and two pose-guided branches, *i.e.*, a foreground-enhanced branch (FEB), and a body part semantics aligned branch (SAB). Occluded-Duke is a larger occluded dataset, which can better reflect the effectiveness of the models. Table 6 shows the results. Model-1 denotes our Baseline, where ResNet50 network is trained followed by ReID loss. We denote whether SAB/FEB is enabled (denoted by On) or not (denoted by Off) in training in the column titled by $\mathcal{S}$ (means Switch). For all these schemes, the global feature f_G of the MB is used for testing.

Effectiveness of SAB. As shown in Table 6, we denote our interaction based optimization in SAB as $\mathcal{I}$, and multi-part contrastive loss (for knowledge distillation) in SAB as $\mathcal{M}$ (see Section 3.3). We denote Model-2 (MB+SAB) as a scheme when we add the SAB without $\mathcal{I}$ and $\mathcal{M}$, where the part semantics aligned feature f_P is followed by ReID loss. Then the SAB plays a role of regularizing the backbone feature learning. We can see that Model-2 outperforms Baseline by 0.6%/1.3% in mAP/Rank-1.

When the interaction based optimization $\mathcal{I}$ of SAB is used, *i.e.*, Model-3 (MB+SAB w/ I), the performance is further improved by

Index (Scheme)	SAB			FEB		R1	mAP
	$\mathcal{S}$	$\mathcal{I}$	$\mathcal{M}$	$\mathcal{S}$	C		
1 (Baseline)	Off	×	×	Off	×	52.7	45.9
2 (MB+SAB)	On	×	×	Off	×	54.0	46.5
3 (MB+SAB w/ I)	On	✓	×	Off	×	56.4	48.1
4 (MB+SAB w/ IM)	On	✓	✓	Off	×	59.4	52.0
5 (MB-SAB+FEB)	On	✓	✓	On	×	61.2	52.1
6 (MB-SAB+FEB w/ C)	On	✓	✓	On	✓	**63.0**	**54.1**

Table 6: Effectiveness of our designs in the proposed PGFL-KD on Occluded-Duke. It consists of a main branch (MB), and two pose-guided branches, *i.e.*, a foreground-enhanced branch (FEB), and a body part semantics aligned branch (SAB). We denote the interaction-based otimization in SAB by $\mathcal{I}$, multi-part contrastive loss in SAB by $\mathcal{M}$, and consistent loss in FEB by C. Note that for all these schemes, the global feature f_G of the MB is used for testing.

Method	Rank-1	mAP
MB + FEB + SAB w $\mathcal{L}_{CL}$	59.1	50.0
MB + FEB + SAB w $\mathcal{L}_{MCL}$	63.0	54.1

Table 7: Effectiveness of using different knowledge distillation losses for the SAB in our PGFL-KD on Occluded-Duke.

1.6%/2.4% in mAP/Rank-1 in comparison with Model-2 (MB+SAB). This demonstrates the effectiveness of our proposed interaction-based training in promoting the semantics alignment for the global feature. In Model-4 (MB+SAB w/ IM), the using of the proposed multi-part contrastive loss ($\mathcal{M}$) explicitly enhances the channel-wise feature alignment of the global feature guided by the local part features of the SAB, which brings additional 3.9%/3.0% gain in mAP/Rank-1.

Effectiveness of FEB. We denote the consistent loss (for knowledge distillation) in FEB as C. On top of Model-4 (MB+SAB w/ IM), when adding the FEB without C, we denote the scheme as Model-5 (MB-SAB), where the foreground-enhanced feature f_E is followed by ReID loss. In this case, the FEB regularizes the feature learning of the backbone network. The performance is significantly improved by 0.1%/1.8% in mAP/Rank-1 over Model-4 (MB+SAB w/ IM). The foreground-enhanced operation in FEB intends to emphasise the features of visible body parts while alleviating the interference of obstructions and background. When we explicitly distilling the knowledge from the FEB to the MB by adding consistent loss (*i.e.*, C enabled), we can see that Model-6 (MB-SAB + FEB w/ C) is much superior than Model-5 that without using consistent loss. Model-6 represents our final scheme PGFL-KD. Thanks to our designs, it outperforms Baseline significantly by **8.2%/10.3%** in mAP/Rank-1.

Effectiveness of Multi-part Constrastive Loss vs. Consistent Loss for the SAB. To distill knowledge from the body part semantics aligned feature f_P in the SAB to the global feature f_G in the MB, we use multi-part constrastive loss ($\mathcal{L}_{MCL}$) for better alignment. Table 7 shows that replacing this contrastive loss with a consistent loss $\mathcal{L}_{CL}$ (similar to Eq.(4) in SAB, there is a 3.9% drop in mAP accuracy.

Influence of Different Hyper-parameters. We study the influence of different hyper-parameters on the performance. Figure 3

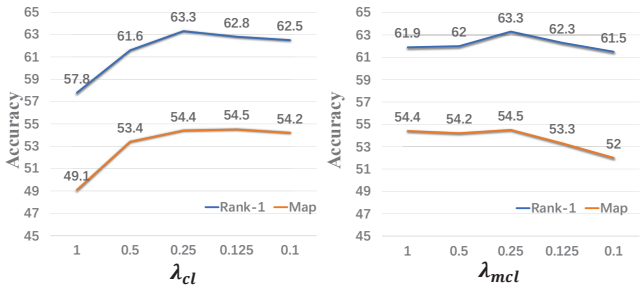

Figure 3: Evaluation of the proposed PGFL-KD with different values of parameter on Occluded-Duke. (a) multi-part contrastive loss λ_{mcl} ;(b) consistent loss λ_{cl}.

Feature for Testing	FLOPS	Param	Rank-1	mAP
Global Feature f_G (Baseline)	8.98G	39.89M	52.7	45.9
Global Feature f_G (PGFL-KD)	8.98G	39.89M	63.0	54.1
Body Part Feature f_P	24.75G	116.09M	57.1	46.7
Body Part Fused Feature f_V	24.75G	117.54M	62.8	53.0
Foreground Feature f_F	24.72G	103.51M	58.3	47.3
Foreground Fused Feature f_E	24.72G	104.95M	63.4	54.6

Table 8: Performance (%) and inference complexity comparisons when we use different features for matching for our PGFL-KD on Occluded-Duke.

shows the results. We can see that when $\lambda_{cl} = 0.25$ and $\lambda_{mcl} = 0.25$, PGFL-KD presents the best performance (in mAP).

4.5 Different Features for Matching and Inference Complexity

For our PGFL-KD scheme, we compare the performance when we use different features for matching in inference, and show the results in Table 8. 1) When we use the global feature f_G in testing, ours significantly outperforms Baseline by 8.3%/10.3% in term of mAP/Rank-1. This is only 0.5%/0.4% inferior to the best performance in term of mAP/Rank-1 which need to use pose information in testing. Through pose-guided interaction learning (*i.e.*, knowledge distillation and interaction-based training), we get rid of the dependency on the pose estimator, retaining high performance and low computational complexity in the test. The computational complexity is the same as Baseline, which is about 1/3 of the schemes which need a pose estimator. 2) Body part feature f_P only or foreground feature f_P only is less effective since it lacks the global information. In contrast, f_G still preserves global information while inheriting the merits of pose-guided features. 3) The ensemble of the features (f_E and f_V) further brings slight gain (about 0.5 in mAP). However, their computational complexity is about three times greater than that of using only the global feature.

We compare the inference speed of our method with PCB [41], the partial re-id methods (DSR [7], and SFR [9]), and the occluded re-id methods PGFA [30]. Table 9 shows that our method is much faster than other methods DSR [7], and SFR [9] because there is no time-consuming feature map matching during inference in our method. Ours has similar inference speed with PGFA w/o pose [30] but achieves much better performance (see Table 2)

Method	Time	Method	Time
DSR [7]	4.84s	SFR [9]	4.65s
PGFA [30]	0.82s	PGFA w/o pose [30]	0.12s
PCB [41]	0.09s	PGFL-KD (Ours)	0.08s

Table 9: Inference speed (seconds per query) on Occluded-Duke.

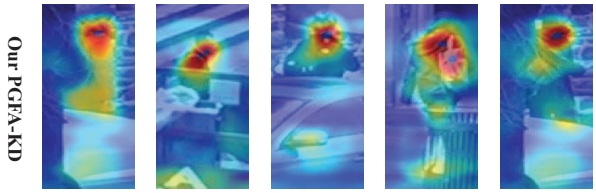

Figure 4: Visualization of the feature corresponds for our PGFL-KD scheme, where the responses for Baseline and the original images are shown in Figure 1.

4.6 Feature Visualization

As discussed in Section 3.2 and 3.3, we expect to let the main branch ignore the interference from obstructions/background and learn semantics aligned representations. We visualize the feature responses F of our PGFL-KD in Figure 4, where the responses for Baseline are shown in Figure 1 (b). In Figure 1 (b), for the regions with objects occluding persons (*i.e.*, obstructions), the networks usually mistakenly generate high responses by regarding them as discriminative person regions. With the guidance of FEB and SAB, the PGFL-KD focuses on the regions more related to foreground objects compared with Baseline.

5 CONCLUSION

In this paper, we propose a network named Pose-Guided Feature Learning with Knowledge Distillation (PGFL-KD). PGFL-KD consists of a main branch (MB), a foreground-enhanced branch (FEB), and a body part semantics aligned branch (SAB). Specifically, the FEB intends to emphasise the features of visible body parts while excluding the interference of obstructions and background (*i.e.*, foreground feature alignment). The SAB encourages different channel groups to focus on different body parts to have body part semantics aligned representation. To get rid of the dependency on pose information and have a model of low complexity when testing, we regularize the main branch to learn the merits of the FEB and SAB through knowledge distillation and interaction-based training. Extensive experiments on occluded, partial, and holistic ReID tasks show the effectiveness of our proposed network and validate the superiority of PGFL-KD over various state-of-the-art methods.

6 ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China under Grand 2020AAA0105702, National Natural Science Foundation of China (NSFC) under Grants U19B2038, the University Synergy Innovation Program of Anhui Province under Grants GXXT-2019-025, and China Postdoctoral Science Foundation Funded Project under Grant 2020M671898.

REFERENCES

[1] Xing Fan, Hao Luo, Xuan Zhang, Lingxiao He, Chi Zhang, and Wei Jiang. 2018. Scpnet: Spatial-channel parallelism network for joint holistic and partial person re-identification. 19–34.

[2] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. 2020. Pose-guided Visible Part Matching for Occluded Person ReID. In *CVPR*. 11744–11752.

[3] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. 2018. FD-GAN: Pose-guided feature distilling GAN for robust person re-identification. 1222–1233.

[4] Jianping Gou, Baosheng Yu, Stephen John Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *IJCV* (2021).

[5] Tengda Han, Weidi Xie, and Andrew Zisserman. 2020. Self-supervised Co-training for Video Representation Learning. 33 (2020).

[6] Yi Hao, Nannan Wang, Jie Li, and Xinbo Gao. 2019. HSME Hypersphere Manifold Embedding for Visible Thermal Person Re-identification. In *AAAI*.

[7] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. 2018. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *CVPR*. 7073–7082.

[8] Lingxiao He and Wu Liu. 2020. Guided Saliency Feature Learning for Person Re-identification in Crowded Scenes. In *ECCV*. 357–373.

[9] Lingxiao He, Zhenan Sun, Yuhao Zhu, and Yunbo Wang. 2018. Recognizing partial biometric patterns. *arXiv preprint arXiv:1810.07399* (2018).

[10] Lingxiao He, Yinggang Wang, Wu Liu, He Zhao, Zhenan Sun, and Jiashi Feng. 2019. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In *ICCV*. 8450–8459.

[11] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In Defense of the Triplet Loss for Person Re-Identification. *arXiv preprint arXiv:1703.07737* (2017).

[12] Bingyu Hu, Jiawei Liu, and Zheng-jun Zha. 2021. Adversarial Disentanglement and Correlation Network for Rgb-Infrared Person Re-Identification. In *ICME*. IEEE, 1–6.

[13] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. 2018. Adversarially occluded samples for person re-identification. In *CVPR*. 5098–5107.

[14] Yukun Huang, Zheng-Jun Zha, Xueyang Fu, Richang Hong, and Liang Li. 2020. Real-world person re-identification via degradation invariance learning. In *CVPR*. 14084–14094.

[15] Sara Iodice and Krystian Mikolajczyk. 2018. Partial Person Re-identification with Alignment and Hallucination. 101–116.

[16] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *ICML*. 448–456.

[17] Xin Jin, Cuiling Lan, Wenjun Zeng, Guoqiang Wei, and Zhibo Chen. 2020. Semantics-aligned representation learning for person re-identification. In *AAAI*.

[18] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. 2018. Human semantic parsing for person re-identification. In *CVPR*. 1062–1071.

[19] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[20] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. 2012. Large scale metric learning from equivalence constraints. In *CVPR*.

[21] Wei Li, Xiatian Zhu, and Shaogang Gong. 2018. Harmonious Attention Network for Person Re-identification. In *CVPR*. 2285–2294.

[22] Shengcai Liao and Stan Z Li. 2015. Efficient psd constrained asymmetric metric learning for person re-identification. In *ICCV*. 3685–3693.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*. 740–755.

[24] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. 2018. Pose Transferrable Person Re-identification. In *CVPR*. 4099–4108.

[25] Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, and Meng Wang. 2019. Adaptive transfer network for cross-domain person re-identification. In *CVPR*. 7202–7211.

[26] Jiawei Liu, Zheng-Jun Zha, Xuejin Chen, Zilei Wang, and Yongdong Zhang. 2019. Dense 3D-convolutional neural network for person re-identification in videos. *TOMM* 15, 1s (2019), 1–19.

[27] Jiawei Liu, Zheng-Jun Zha, QI Tian, Dong Liu, Ting Yao, Qiang Ling, and Tao Mei. 2016. Multi-scale triplet cnn for person re-identification. In *ACM MM*. 192–196.

[28] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. 2019. Bag of tricks and a strong baseline for deep person re-identification. 0–0.

[29] Hao Luo, Wei Jiang, Xing Fan, and Chi Zhang. 2020. Stnreid: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification. *IEEE Transactions on Multimedia* (2020).

[30] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. 2019. Pose-Guided Feature Alignment for Occluded Person Re-Identification. In *ICCV*.

[31] Hyunjong Park and Bumsub Ham. 2020. Relation Network for Person Re-identification. In *AAAI*, Vol. 34. 11839–11847.

[32] Lei Qi, Jing Huo, Lei Wang, Yinghuan Shi, and Yang Gao. 2018. Maskreid: A mask based deep ranking neural network for person re-identification. *arXiv preprint arXiv:1804.03864* (2018).

[33] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*. 17–35.

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. In *ICML*, Vol. 115. 211–252.

[35] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. 2018. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *CVPR*. 420–429.

[36] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. 2018. Mask-guided contrastive attention model for person re-identification. In *CVPR*. 1179–1188.

[37] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. 2017. Pose-Driven Deep Convolutional Model for Person Re-identification. In *ICCV*. 3980–3989.

[38] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. 2018. Part-Aligned Bilinear Representations for Person Re-identification. In *ECCV*. 418–437.

[39] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. In *CVPR*.

[40] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. 2019. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *CVPR*. 393–402.

[41] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*. 480–496.

[42] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. 2020. High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-Identification. In *CVPR*. 6449–6458.

[43] Wen Wang, Yang Cao, Jing Zhang, Fengxiang He, Zheng-Jun Zha, Yonggang Wen, and Dacheng Tao. 2021. Exploring Sequence Feature Alignment for Domain Adaptive Detection Transformers. *arXiv preprint arXiv:2107.12636* (2021).

[44] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. 2021. Deep learning for person re-identification: A survey and outlook. *TPAMI* (2021).

[45] Qian Yu, Xiaobin Chang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. 2017. The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching. *arXiv preprint arXiv:1711.08106* (2017).

[46] Wei Zhai, Yang Cao, Zheng-Jun Zha, HaiYong Xie, and Feng Wu. 2020. Deep structure-revealed network for texture recognition. In *CVPR*. 11010–11019.

[47] Wei Zhai, Yang Cao, Jing Zhang, and Zheng-Jun Zha. 2019. Deep multiple-attribute-perceived network for real-world texture recognition. In *ICCV*. 3613–3622.

[48] Jing Zhang, Zhe Chen, and Dacheng Tao. 2021. Towards high performance human keypoint detection. *International Journal of Computer Vision* (2021), 1–24.

[49] Jing Zhang and Dacheng Tao. 2020. Empowering things with intelligence: a survey of the progress, challenges, and opportunities in artificial intelligence of things. *IEEE Internet of Things Journal* 8, 10 (2020), 7789–7817.

[50] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. 2019. Densely Semantically Aligned Person Re-Identification. In *CVPR*.

[51] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. 2017. Deeply-Learned Part-Aligned Representations for Person Re-identification. In *ICCV*. 3239–3248.

[52] Kecheng Zheng, Cuiling Lan, Wenjun Zeng, Zhizheng Zhang, and Zheng-Jun Zha. 2021. Exploiting Sample Uncertainty for Domain Adaptive Person Re-Identification. In *AAAI*, Vol. 35. 3538–3546.

[53] Kecheng Zheng, Wu Liu, Lingxiao He, Tao Mei, Jiebo Luo, and Zheng-Jun Zha. 2021. Group-aware Label Transfer for Domain Adaptive Person Re-identification. In *CVPR*.

[54] Kecheng Zheng, Wu Liu, Jiawei Liu, Zheng-Jun Zha, and Tao Mei. 2020. Hierarchical Gumbel Attention Network for Text-based Person Search. In *ACM MM*. 3441–3449.

[55] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *ICCV*. 1116–1124.

[56] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. 2015. Partial Person Re-Identification. In *ICCV*. 4678–4686.

[57] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2017. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896* (2017).

[58] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *ICCV*. 2242–2251.

[59] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. 2020. Identity-Guided Human Semantic Parsing for Person Re-Identification. In *ECCV*.

[60] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. 2018. Occluded person re-identification. In *ICME*. 1–6.

[61] Jiaxuan Zhuo, Jianhuang Lai, and Peijia Chen. 2019. A Novel Teacher-Student Learning Framework For Occluded Person Re-Identification. *arXiv preprint arXiv:1907.03253* (2019).