Occlude Them All: Occlusion-Aware Attention Network for Occluded Person Re-ID

Peixian Chen[1,3], Wenfeng Liu[1], Pingyang Dai[1]*, Jianzhuang Liu[2],
Qixiang Ye[4], Mingliang Xu[5], Qi'an chen[1], Rongrong Ji[1]
[1]Xiamen University, China, [2]Noah's Ark Lab, Huawei Technologies, [3]Tencent YouTu Lab
[4]University of Chinese Academy of Sciences, China, [5]Zhengzhou University, China,
pxchen@stu.xmu.edu.cn, wenfengliu.xmu.edu.cn, pydai@xmu.edu.cn, liu.jianzhuang@huawei.com,
qxye@ucas.ac.cn, iexumingliang@zzu.edu.cn, cheer@xmu.edu.cn, rrji@xmu.edu.cn,

Abstract

Person Re-Identification (ReID) has achieved remarkable performance along with the deep learning era. However, most approaches carry out ReID only based upon holistic pedestrian regions. In contrast, real-world scenarios involve occluded pedestrians, which provide partial visual appearances and destroy the ReID accuracy. A common strategy is to locate visible body parts by auxiliary model, which however suffers from significant domain gaps and data bias issues. To avoid such problematic models in occluded person ReID, we propose the Occlusion-Aware Mask Network (OAMN). In particular, we incorporate an attention-guided mask module, which requires guidance from labeled occlusion data. To this end, we propose a novel occlusion augmentation scheme that produces diverse and precisely labeled occlusion for any holistic dataset. The proposed scheme suits real-world scenarios better than existing schemes, which consider only limited types of occlusions. We also offer a novel occlusion unification scheme to tackle ambiguity information at the test phase. The above three components enable existing attention mechanisms to precisely capture body parts regardless of the occlusion. Comprehensive experiments on a variety of person ReID benchmarks demonstrate the superiority of OAMN over state-of-the-arts.

1. Introduction

Person Re-Identification (ReID) aims to identify the same pedestrian captured by different cameras under varying viewpoints, lights, and locations. Along with the deep learning era, ReID approaches based on Convolution Neural Networks (CNNs) have achieved remarkable performance [2, 32, 14, 12]. However, these approaches carry out ReID only based upon the *holistic* pedestrian regions, which ignore the *occlusion* that happens frequently in real-world scenarios as shown in Figure 1a.

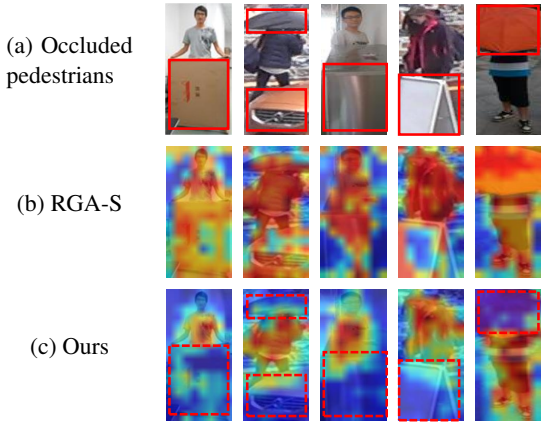

Figure 1: Examples of occluded pedestrians and illustrations of introducing attentions. (a) shows examples of occluded pedestrians. (b) and (c) illustrate the attention introduced by RGA-S [33] and our proposed OAMN. (b) may erroneously focus on the occlusion, but (c) does not.

Identifying occluded pedestrians faces essential challenges. In particular, the occluded pedestrian contains fewer distinguishable features from the pedestrian itself, while introducing ambiguity information from the occluded regions. Such ambiguity, like rich texture and noise, misleads the appearance representation. Existing approaches typically employ auxiliary models to obtain information for occluded body parts to assist the learning procedure, such as capturing body-part features with Human parsing [10], separating

*Corresponding author.

(a) Real-world Occlusions (b) Occlusion Augmentation

Figure 2: Occluded pedestrians and augmented images.

occlusions with Body Mask [42], and locating human regions by estimating the person's key points [19, 30, 28]. However, such auxiliary models are pre-trained on different domains, making the learning procedure suffer from significant domain gap and data bias issues [8]. Hence, incorrect labels mislead the learning procedure, while the entire network becomes more complicated and inefficient [17].

To avoid the auxiliary model's domain gap and inefficiency issues, we propose an *attention-guided mask module* that incorporates the attention mechanism, which has been shown effective in guiding the model to focus on body regions instead of backgrounds [29, 33, 13]. However, such benefits cannot directly transfer to the occlusion problem, as real-world occlusions contain richer textural information than the backgrounds. We have observed erroneous results when directly using the attention as shown in Figure 1b. Although models learned with cutting-edge attention networks (*i.e.*, RGA-S [33]) are able to avoid background objects, such models cannot disambiguate occlusions effectively. One common cause of this problem is the lack of guidance from labeled occlusion data.

To supplement the labeled occlusion data, we propose a novel *occlusion augmentation* scheme that produces diverse occluded images with more precise auxiliary labels. Empirically, in various real-world scenarios, common occlusions can be categorized into four locations (top, bottom, left, right) and two areas (half, quarter), as being exampled in Figure 2a. Correspondingly, the proposed scheme augments training data with the above occlusion types using "occlusion" cropped from other images and inherits the original training image's label, as shown in Figure 2b. Hence, the proposed scheme is more suitable to real-world occlusions than existing data augmentation schemes, which utilize random erasing [40] or random occluding [42] techniques that consider only monotonous occlusions. Models learned with such augmented data can easily overfit to a particular set of occlusions, providing limited improvements. In contrast, the proposed scheme suits better to real-world scenarios through diverse occluded images and precise labels.

Still, test images come with diverse and unlabeled occlusions at the test phase, destroying the ReID performance. We propose an *occlusion unification* scheme to tackle this problem. First, we label the target pedestrian's occlusion type by learning an auxiliary occlusion grader. Second, we mitigate the diversity by occluding all gallery images with the same type of occlusion as the target pedestrian, namely the "occlude them all" strategy (with few exceptions as detailed in Section 3.4). Hence, the original ambiguity information is unified, allowing the attention module to precisely capture body parts regardless of the occlusion.

In summary, we propose an Occlusion-Aware Mask Network (OAMN) to address the occlusion problem in person ReID. OAMN employs three innovative components: the attention-guided mask module, occlusion augmentation, and occlusion unification. These components enable existing attention mechanisms to precisely capture body parts regardless of the occlusion, as shown in Figure 1c. OAMN tackles several challenges to finally bring attention mechanisms to occluded person ReID.

We summarize our contributions as follows.

1. We propose the Occlusion-Aware Mask Network, an efficient and effective approach to address the occlusion problem in person ReID. We enable attention mechanisms to precisely capture body parts regardless of the occlusion.

2. We propose a new occlusion augmentation scheme to produce diverse occluded images and precise labels for any holistic datasets. We propose a novel occlusion unification scheme to unify ambiguity at the test phase.

3. We evaluate the proposed OAMN in three person ReID datasets containing occlusions. Quantitative results show that OAMN achieves state-of-the-art performance, with the rank-1 accuracy of 62.6%, 86.0%, and 77.3% on Occluded-DukeMTMC, Partial-ReID, and Partial-iLIDS, respectively.

2. Related Work

2.1. Person Re-Identification

Person re-identification aims to spot a person of interest in other cameras and great progress of this research has been made in recent years. Instead of hand-crafted descriptors [32, 18] and metric learning methods [1, 6, 39], deep learning algorithms [27, 33, 13] have become dominant in person re-identification nowadays. Some methods attempt to learn the local information to achieve finer-grained feature matching [27, 31, 34, 35]. The attention mechanism has also been adopted to ensure the model to focus on human areas, which results in more effective features [29, 33, 13]. However, these methods ignore the occlusion problem and they cannot separate the person from the occlusion, which is inevitable in the real worlds especially in the crowd scenes.

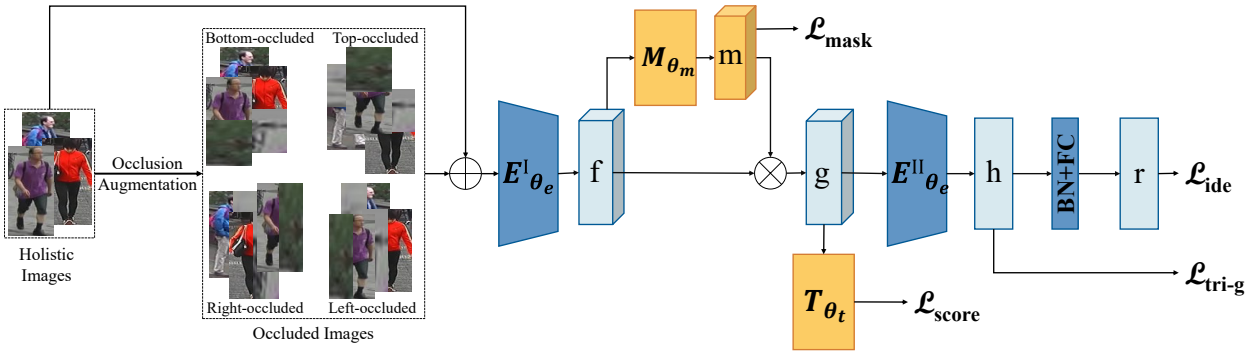

Figure 3: An overview of OAMN. An augmentation mechanism produces additional occluded images. We split the backbone encoder into the lower part E^{I} and higher part E^{II}. Between the split encoder, we insert an attention-guided mask network M to guide the network's attention. An occlusion grader T learns to predict the occlusion type at the test phase.

2.2. Partial Person Re-Identification

Partial person ReID aims to manually crop the occluded probe's visible part as the new probe image and then match the partial probe image to gallery holistic images. Zheng *et al.* [37] first proposed the partial ReID problem. They adopt a model named Ambiguity-sensitive Matching Classifier (AMC) to match the global-to-local information and introduce a global part-based matching model called Sliding Window Matching (SWM). Sun *et al.* [26] propose a Visibility-aware Part Model (VPM) for the partial person ReID task to locate visible regions on pedestrians' images through self-supervision. He *et al.* propose the Deep Spatial feature Reconstruction (DSR) [6] for partial person re-id, which is alignment-free and flexible to arbitrary-sized person images. Luo *et al.* present a novel deep partial ReID framework based on pairwise Spatial Transformer Networks (STNReID) [16], which can be trained on holistic datasets. However, the partial person Re-ID problem needs a manually cropping operation, which is inefficient and might introduce human bias. Even though there has been enormous progress in partial person ReID, it is still not enough to overcome the occlusion problem.

2.3. Occluded Person Re-Identification

The study of the occlusion in person ReID proposed by Zhou *et al.* [41] is different from the partial person ReID. It directly extracts the features form the occluded person images without cropping away the invisible part, which is more practical in real-world scenarios than the partial person ReID. Recent study methods in this topic can be divided into two categories: semantic segmentation [42, 8, 10] and posed guidance with human key-points [19, 30, 3].

Semantic segmentation is used in some works. For example, Zhuo *et al.* [42] train a co-saliency branch, in which the ground truth comes from the masks predicted by an existing salient object detector. He *et al.* propose FPR [8], an alignment-free approach using semantic segmentation models [21] to obtain the person mask reconstruction. Huang *et al.* [10] adopt human parsing to address the problem.

As for the pose-guided methods with human key-points, Miao *et al.* introduce a method named Pose-Guided Feature Alignment (PGFA) [19], exploiting pose landmarks to disentangle useful information from occlusion noise. Gao *et al.* propose PVPM [3] that jointly learns discriminative features and pose-guided attention to obtain useful information by graph matching. Wang *et al.* [30] utilize human key-points to extract local features and predict similarity scores using topological information. However, these methods still utilize a pre-trained model, which introduces data bias that limits the performance while making the network complex.

Differing from all the above methods, our approach does not rely on extra models. Inspired by [33], we combine the attention approach with our new data augmentation method that is different from [40] to solve the hard occlusion problem in person ReID. In this simple way, we save much time in the test phase and achieve great performance. Moreover, our method can also help to enlarge the occlusion datasets which are still limited so far.

3. Method

In this section, we introduce the proposed Occlusion-Aware Mask Network (OAMN). Figure 3 demonstrates the overview of OAMN. It begins with an occlusion augmentation mechanism that produces training images. We split the backbone encoder $E(\cdot)$ into two parts *i.e.*, $E^{\mathrm{I}}(\cdot)$ and $E^{\mathrm{II}}(\cdot)$, so that we can insert an attention module between them. The backbone encoder is parameterized by θ_e.

For each input image x, the lower-part encoder $E^{\mathrm{I}}(\cdot)$ maps the input to a feature $f \triangleq E^{\mathrm{I}}(x)$. The following attention-guided mask module $M(\cdot)$ parameterized by θ_m produces a spatial attention map $m \triangleq M(f)$, which is used

to generate the attentive feature $g \triangleq m \odot f$ by element-wise multiplications. We utilize the attentive feature g in two ways. First, it forwards the remaining higher-part encoder $E^{\text{II}}(\cdot)$ to obtain the final representation h and classification logits r. Second, an auxiliary grader $T(\cdot)$ parameterized by θ_t predicts the input feature's occlusion type.

3.1. Occlusion Augmentation

Existing person ReID approaches fail to handle occluded person images. One factor that limits their robustness against occlusion is the lack of occlusion data. As such, we propose the occlusion augmentation, a novel scheme that produces *diverse* and *labeled* occlusion data.

Empirically, in various real-world scenarios, common occlusions can be coarsely categorized into four locations (top, bottom, left, right) and two areas (half, quarter). Rare cases are ignored, where more than half of the pedestrian is occluded. Correspondingly, our proposed scheme augments training data with the above occlusion types by following a three-step process: (1) Randomly choose a training image x, from which we crop a rectangular patch p; (2) Randomly scale the patch p to one of the two areas w.r.t. the input image; (3) Put the scaled patch onto each of the four locations of the input image, respectively.

Since we regard cropped patches as the occlusion, one critical design is to avoid cropping human bodies. To this end, we crop the patch from the corners of the chosen image. We formally describe the above process as follows. *Firstly*, letting a' denote the area of a chosen image x, we determine the cropped patch size $p_h \times p_w$ such that $p_h^2 = a \cdot r$ and $p_w^2 = a/r$, where $a \triangleq \epsilon \cdot a'$ is the reduced area with $\epsilon \sim \mathcal{U}(0.02, 0.2)$ and $r \sim \mathcal{U}(0.3, 3.3)$ is the ratio of p_h and p_w. $\mathcal{U}$ denotes the uniform distribution. *Secondly*, we choose the cropped patch's location at random from four corners $(x, y) \in \{0, h - p_h\} \times \{0, w - p_w\}$, where $h \times w$ denotes the chosen image x's size. *Thirdly*, we scale and put the obtained patch onto the target image described above. Following this process, we obtain four occluded copies of each training image. We denote these copies by $x^{\mathtt{p}}$, where $\mathtt{p} \in \mathcal{P} \triangleq \{\mathtt{t}, \mathtt{b}, \mathtt{l}, \mathtt{r}\}$ refers to the occlusion's location: top, bottom, left or right. Particularly, we omit the superscript $\mathtt{p}$, or set $\mathtt{p} = \backslash$, to denote the original holistic image x, or non-occlusion. As such, $x^{\mathtt{p}}$ inherits the same label from x.

3.2. Attention-Guided Mask Module

Given the above occlusion augmentation scheme, we are able to guide the model to learn non-occluded body parts using the attention mechanism. In specific, we generate a spatial weight map using RGA-S [33]. Since we observe that *textural* features could mislead the network's attention, we exploit *intermediate-layer* features to capture the contour information.

Extending from the above analysis, we propose an attention-guided mask module to generate spatial weight maps m for each input feature, formally described as:

$$\begin{aligned} m^{\mathtt{p}} &= M(f^{\mathtt{p}}), \\ g^{\mathtt{p}} &= m^{\mathtt{p}} \odot f^{\mathtt{p}}, \end{aligned} \tag{1}$$

where $\odot$ is the element-wise multiplication, f and g denote the input and output features, respectively. Below, we derive two constraints to guide the learning of masked features.

First, we expect the network to capture human features as complete as possible, even when given occluded data. Due to our augmentation, the body parts remaining in the occluded image are identical to the holistic image's corresponding parts. Hence, the attention mask learned from occluded images should ideally focus on the same area if applied to occluded and holistic images, respectively. This constraint also prevents attention masks from erroneously focusing on the augmented occlusion. Therefore, we minimize the ℓ_2 distance between the attentive part of occluded and holistic features:

$$\mathcal{L}_{\text{mask1}} = \frac{1}{n} \sum_{\mathtt{p} \in \mathcal{P}} \sum_{i=1}^{n} \left\| (f_i - f_i^{\mathtt{p}}) \odot m^{\mathtt{p}} \right\|_2^2, \tag{2}$$

where n denotes the batch size.

Second, features of symmetrically occluded images (*i.e.*, top vs. bottom and left vs. right) should ideally capture the complete information when combined. For example, features from images that are top-half and bottom-half occluded, when combined together, should recover the complete feature. Therefore, we minimize their ℓ_2 distances to the complete feature:

$$\mathcal{L}_{\text{mask2}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left(\left\| g_i^{\mathtt{t}} + g_i^{\mathtt{b}} - g_i \right\|_2^2 + \left\| g_i^{\mathtt{l}} + g_i^{\mathtt{r}} - g_i \right\|_2^2 \right), \tag{3}$$

where $\mathcal{I}$ denotes the set of indices of half-occluded features (we augment each training image with half/quarter occlusion area at random). Note that all $g_i^{\mathtt{p}}$'s are obtained from the i-th training image but with different occlusion types.

In summary, we impose the following constraint to guide the proposed mask module:

$$\mathcal{L}_{\text{mask}} = \alpha_1 \cdot \mathcal{L}_{\text{mask1}} + \alpha_2 \cdot \mathcal{L}_{\text{mask2}}, \tag{4}$$

where α_1 and α_2 are two hyper-parameters that control the trade-off between the two terms.

3.3. Grouped Triplet Loss

The triplet loss is another effective criterion to assist similarity learning in person ReID [9]. In particular, it reduces intra-class distances while enlarging inter-class ones. Typically, the triplet loss function $\mathcal{L}_{\text{tri}}(\cdot)$ is defined as

$$\mathcal{L}_{\text{tri}}(h_{\text{a}}, h_{\text{p}}, h_{\text{n}}) = \left(\|h_{\text{a}} - h_{\text{p}}\|^2 - \|h_{\text{a}} - h_{\text{n}}\|^2 + m_{\text{tri}} \right)_+, \tag{5}$$

where h is the representation, m_{tri} is the margin, subscripts a, p, and n are the anchor, positive, and negative samples.

However, this common triplet loss is incompatible with our approach. Due to the randomness in our proposed occlusion augmentation (see Section 3.1), every data batch contains mixed types of occlusions. Blindly reducing the distance between occluded and non-occluded features may have negative impacts on learned attention areas. As such, there is a need for an occlusion-aware triplet loss. Hence, we propose the grouped triplet loss $\mathcal{L}_{\text{tri-g}}$ to handle different types of occlusion separately. It is defined as

$$\mathcal{L}_{\text{tri-g}} = \frac{1}{|\mathcal{P}'|} \sum_{\mathtt{p} \in \mathcal{P}'} \sum_{i=1}^{n} \mathcal{L}_{\text{tri}}(h_i^{\mathtt{p}}, h_{i\text{-}\mathtt{p}}^{\mathtt{p}}, h_{i\text{-}\mathtt{n}}^{\mathtt{p}}), \tag{6}$$

where $\mathcal{P}' = \mathcal{P} \cup \{\backslash\}$ is the set of all four occlusion locations with the non-occlusion case (denoted by "\"). For each type of occluded features, we search for the positive and negative samples with the corresponding occlusion type. We define the positive/negative sample as the farthest/nearest sample with the same/different label as the anchor.

3.4. Occlusion Unification

At the test phase, some test images come with diverse and unlabeled occlusions, which destroy the ReID performance. We propose an occlusion unification scheme to tackle this problem. Differing from the training phase that desires diversity, the test phase avoids diversity to reduce the ambiguity. As such, we mitigate the occlusion's diversity by the "occlude them all" strategy. However, such unification relies on the knowledge of occlusion types, which are typically unknown in the test stage.

To this end, we utilize the augmented occlusion data and learn a supervised grader, which identifies the input image's occlusion type at the test phase. In Section 3.1, we consider occlusions of four locations and two areas, forming in total eight occlusion types. However, simply employing an eight-class classifier is problematic. To clarify, although such occlusion types can supplement the augmentation, they are not precise enough to quantify diverse occlusions in the real-world. Thus, deploying such a classifier might overfit the limited occlusion types that we have considered for augmentation. To circumvent this problem, we propose a threshold-based occlusion grader. The occlusion grader $T(\cdot)$ consists of a fully-connected layer and a softmax operator. It outputs a four-dimension score vector $s = \{s^{\mathtt{t}}, s^{\mathtt{b}}, s^{\mathtt{l}}, s^{\mathtt{r}}\}$, indicating the score of occlusions at four locations. We interpret such a score as the occlusion's area. At the training stage, we define the score s of augmented data $x^{\mathtt{p}}$ such that $s^{\mathtt{p}} = 2 \cdot a_{\text{aug}}$, where a_{aug} denotes the occlusion's area we augmented (i.e., $1/2$ or $1/4$). We learn this grader with the following loss function:

$$\mathcal{L}_{\text{score}} = \frac{1}{n \cdot |\mathcal{P}'|} \sum_{\mathtt{p} \in \mathcal{P}'} \sum_{i=1}^{n} \|T(g_i^{\mathtt{p}}) - s_i\|_2^2, \tag{7}$$

where g_i is the masked feature and s_i is the corresponding occlusion score. Note that this grader is independent of other modules. We do not back propagate its gradients to other modules like the encoder and mask network.

At the test stage, we employ a threshold-based method to deduce the occlusion type. In particular, for a given score s, we consider $\mathtt{p} = \arg\max s$ as the occluded location and $s^{\mathtt{p}}$ as the occluded area. We reduce the occlusion area a_{q} of query images to three cases:

$$a_{\mathrm{q}} = \begin{cases} 0, & \text{if } s^{\mathtt{p}} < 0.5, \\ 1/4, & \text{if } 0.5 \le s^{\mathtt{p}} < 0.75, \\ 1/2, & \text{if } s^{\mathtt{p}} \ge 0.75, \end{cases} \tag{8}$$

which determines the occlusion's location $\mathtt{p}$ and area a_{q}. We then employ the "occluded them all" strategy, which occludes all gallery images with the same occlusion as the target pedestrian by masking out the occlusion regions. We may also occlude the query image correspondingly if the gallery image already contains occlusion. Hence, the original ambiguity information is unified, allowing us to compare identically-occluded gallery features $\mathcal{G} \triangleq \{h_i^{\mathtt{p}}\}_{i=1}^{n_{\mathrm{g}}}$ and the query feature $h_q^{\mathtt{p}}$, where n_{g} is the size of gallery set.

3.5. Overall Objective Function

In this subsection, we explain the identity objective and summarize the overall objective function. Similar to Section 3.3, we operate on all five occlusion locations (including the non-occluded case). Thus, the identity loss $\mathcal{L}_{\text{ide}}$ can be written as

$$\mathcal{L}_{\text{ide}} = \frac{1}{n \cdot |\mathcal{P}'|} \sum_{\mathtt{p} \in \mathcal{P}'} \sum_{i=1}^{n} \ell_{\text{CE}}(r_i^{\mathtt{p}}, y_i), \tag{9}$$

where (r_i, y_i) denotes the final logits and ground truth label of the i-th input within a mini-batch.

In summary, we obtain the following loss functions

$$\begin{aligned} \mathcal{L}_1 &= \lambda_1 \cdot \mathcal{L}_{\text{mask}} + \lambda_2 \cdot \mathcal{L}_{\text{tri-g}} + \lambda_3 \cdot \mathcal{L}_{\text{ide}}, \\ \mathcal{L}_2 &= \lambda_4 \cdot \mathcal{L}_{\text{score}}, \end{aligned} \tag{10}$$

where λ_1, λ_2, λ_3 and λ_4 are hyper-parameters that control the trade-off between different losses. We minimize $\mathcal{L}_1$ over the entire network and $\mathcal{L}_2$ only over the occlusion grader.

4. Experiments

We conduct experiments on three occluded/partial person ReID benchmarks and two holistic datasets to evaluate the performance of our approach.

Methods	Type	Occluded-Duke			Partial-REID			Partial-iLIDS		
		Rank-1	Rank-5	*m*AP	Rank-1	Rank-5	*m*AP	Rank-1	Rank-5	*m*AP
IDE [5]	H	39.4	57	27.8	57.0	76.3	53.6	68.9	84.9	72.4
PCB [27]	H	42.6	57.1	33.7	66.3	84.0	63.8	46.8	-	40.2
Random-Erasing [40]	H	40.5	59.6	30.0	54.3	75.0	54.4	68.1	82.4	75.1
RGA-S [33]	H	47.3	64.0	38.5	62.0	79.3	59.1	75.6	89.1	78.6
FD-GAN [4]	H	40.8	-	-	-	-	-	-	-	-
DSR [6]	P	40.8	58.2	30.4	73.7	-	68.07	64.3	-	58.1
SFR [7]	P	42.3	60.3	32	56.9	-	-	63.9	-	-
TCSDO [42]	O	-	-	-	82.7	91.3	85.57	-	-	-
FPR [21]	O	-	-	-	81.0	-	76.6	68.1	-	61.8
PGFA [19]	O	51.4	68.6	37.3	69.0	84.7	61.5	71.4	85.7	74.7
PVPM+Aug [3]	O	-	-	-	78.3	89.7	72.3	-	-	-
HOReID [30]	O	55.1	-	43.8	85.3	-	-	72.6	-	-
OAMN (Ours)	O	**62.6**	**77.5**	**46.1**	**86.0**	**91.7**	**77.4**	**77.3**	**86.6**	**79.5**

Table 1: Comparison with state-of-the-art methods on different datasets: Occluded-Duke [19], Partial-REID [37], and Partial-iLIDS [6]. The method types include Holistic, Partial, and Occluded.

4.1. Datasets and Evaluation Setting

Occluded-DukeMTMC [19] consists of 15,618 training images of 702 persons, 2,210 query images of 519 persons, and 17,661 gallery images of 1,110 persons. It is the most challenging one due to the large ratio of occluded persons.

Partial-REID [37] contains 600 images collected from 60 persons under different viewpoints, background, and occlusions. The gallery set contains only holistic images while the query set contains only occluded images.

Partial-iLIDS [6] contains 238 images from 119 persons, captured in the airport where people are typically occluded by luggage or other people. All probes are occluded person images but all gallery images are holistic.

Market-1501 [36] is a common holistic dataset. It contains 12,936 training images of 751 persons, 19,732 query images and 3,368 gallery images of 750 persons captured from 6 cameras. There is only few occluded images.

DukeMTMC-reID [38] contains 16,522 training images of 702 persons, 2,228 queries of 702 persons, and 17,661 gallery images of 702 persons. It is regarded as holistic as it contains much more holistic images than occluded ones, so that this dataset can be treated as a holistic re-id dataset.

Evaluation Protocol. To perform a fair comparison with existing methods, all experiments follow the common evaluation settings in person ReID methods. The Cumulative Matching Characteristic (CMC) and mean Average Precision (*m*AP) are adopted to evaluate the performance. All experiments are performed in the single query setting.

Evaluation Settings. We use ResNet50 [5] and RGA-S [33] as the backbone of our encoder and attention-guided mask module, respectively. The output of ResNet50's layer-3 is set to the input of Attention-Guide Mask Module. All input images are resized to 256×128. We train our network in an end-to-end fashion using the SGD optimizer [22] and batch size 64, which contains 32 identities and 2 examples per identity. We conduct all experiments on three RTX 2080Ti GPUs. The triplet loss's margin m_{tri} in eq. (5) is 0.5. In eq. (4), α_1 and α_2 are set to 1.0. In eq. (10), λ_1, λ_2, λ_3 and λ_4 are set to 5.0, 0.5, 1.0 and 1.0, respectively.

4.2. Comparison with State-of-the-art Methods

We compare our OAMN with existing state-of-the-art (SOTA) person ReID approaches on three different types of datasets, including occluded, partial, and holistic dataset.

4.2.1 Comparisons on Occluded Datasets

The results on Occluded-DuckMTMC are shown in Table 1. The proposed OAMN outperforms SOTAs by large margin with the Rank-1 accuracy of 62.6% and *m*AP of 46.1%.

Holistic Methods. Existing holistic methods typically employ several techniques to improve performance [27, 40, 33, 4]. However, all these methods fail to obtain significant performance gains on occluded datasets due to the lack of occlusion information. Specifically, OAMN significantly improve the Rank-1 accuracy by 15.3% and *m*AP by 7.6% over RGA-S. These result demonstrate that the use of attention mechanism points to erroneous attention at the occlusion.

Partial Methods. We observe that existing partial person ReID approaches such as DSR [6] and SFR [7], still fail to achieve satisfactory performance on occluded datasets. When compared to SFR, our performance improvements are significant, *e.g.*, the boost in Rank-1 and *m*AP is 20.3% and 14.1%. These results demonstrate that, while partial ReID shares a similar challenge as the occluded ReID, such methods are not effective in addressing the occlusion.

Occluded Methods. We compare with SOTA methods PGFA [19], PVPM [3], and HOReID [30] for occluded person ReID. Although these SOTAs use key-point models to assist the model training, the performance gains increase model complexity. Despite this, our OAMN improves the

Rank-1 accuracy by 7.5% and *m*AP by 2.3% over the current best method, *e.g.* HOReID, for occluded person ReID.

4.2.2 Comparisons on Partial Datasets

The measured performance on partial datasets is reported in the last two columns of Table 1. We follow the common training protocol in existing partial person ReID methods [42, 8, 10, 19, 30, 28]. Specifically, we use Market-1501 as the training set and the two partial datasets (Partial-REID and Partial-iLIDS) as the test set. We compare with partial methods [6, 7], occluded methods with a segmentation network [42, 21], and occluded methods with keypoint models [19, 30, 3]. Similar to the results on occluded datasets, OAMN consistently outperform current SOTAs: it improves HOReID's Rank-1 accuracy by 0.7% on Partial-REID and 4.7% on Partial-iLIDS.

4.2.3 Comparisons on Holistic Datasets

The performance on holistic datasets is reported in Table 2. We achieve 93.2% and 86.3% accuracy on Market-1501 and DukeMTMC-reID datasets. We outperform a variety of cutting-edge holistic methods, such as PCB [27], VPM [26], DuATM [23], SPReID [11], MaskReID [20], MGCAM [24], PDC [25], and Pose-transfer [15]. We also outperform the occluded method PGFA [19]. Even when compared with occluded methods that use stronger baseline models, such as FPR [21] and HOReID [30], we can obtain competitive results. Moreover, our approach boosts the commonly used baseline model by a larger margin of 7.5%. In contrast, other occluded methods can only improve their corresponding baselines by less than 3%. Such results show that our method does not overfit to augmented occlusions, hence performing well on holistic datasets.

4.3. Module Performance

In this section, we conduct detailed experiments to study the performance of each module, including occlusion augmentation, grouped triplet loss, and mask module.

4.3.1 Occlusion Augmentation

The proposed scheme augments training data with occlusion at four locations and two areas. We validate the design of this scheme by varying the occlusion's location and area. We present the results in Table 3.

Location. Partial-REID contains occluded images with no significant difference between the number of images with different occlusion locations. As such, we observe a significant performance degradation if removing some types of occlusion. The other two datasets contain more bottom-occluded images. We observe more degradation only when disabling the top and bottom occlusion augmentation.

Methods	Market-1501		DukeMTMC-reID	
	Rank-1	*m*AP	Rank-1	*m*AP
PCB [27]	92.3	77.4	81.8	66.1
VPM [26]	93.0	80.8	83.6	72.6
DuATM [23]	91.4	76.6	-	-
SPReID [11]	92.5	81.3	-	-
MaskReID [20]	90.0	75.3	-	-
MGCAM [24]	83.8	74.3	46.7	46.0
PDC [25]	84.2	63.4	-	-
Pose-transfer [15]	87.7	68.9	30.1	28.3
PGFA [19]	91.2	76.8	82.6	65.5
FPR [21] Baseline	94.1	84.6	87.3	76.2
FPR [21]	95.4	86.6	88.6	78.4
HOReID [30] Baseline	92.6	77.7	83.8	69.7
HOReID [30]	94.2	84.9	86.9	75.6
OAMN (Ours) Baseline	85.7	66.1	80.1	61.6
OAMN (Ours)	93.2	79.8	86.3	72.6

Table 2: Comparison with state-of-the-art methods on holistic datasets. Dashed lines separate methods that use different and stronger baselines. OAMN outperforms most methods by a large margin. Even FPR and HOReID have used different and stronger baselines, OAMN obtains significantly more improvements to the baseline.

Occlusion Type		Occluded-Duke		Partial-REID		Partial-iLIDS	
Location	Area	Rank-1	*m*AP	Rank-1	*m*AP	Rank-1	*m*AP
{\, t, b}	{1/2, 1/4}	56.5	41.0	73.3	67.3	71.1	74.8
{\, l, r}	{1/2, 1/4}	48.0	36.6	74.0	65.7	66.4	69.6
{\, t, b, l, r}	{1/2}	60.1	45.3	83.7	76.7	68.9	72.5
{\, t, b, l, r}	{1/4}	52.8	39.6	72.0	65.4	76.5	79.4
{\, t, b, l, r}	{1/2, 1/4}	**62.6**	**46.1**	**86.0**	**77.4**	**77.3**	**79.5**

Table 3: Comparing different occlusions types.

Area. Occluded-Duke and Partial-REID mainly contain half-occluded images. We observe more degradation when we only augment quarter occlusions. In contrast, datasets with smaller-than-half occlusions show less degradation.

Summary. We empirically observe that all occlusion types we considered are effective. They obtain the best results on all datasets if all types of augmentation are enabled.

4.3.2 Grouped Triplet Loss

We propose the grouped triplet loss to explicitly handle different types of occluded images. In this section, we study its effectiveness by comparing it with the vanilla triplet loss.

As our augmented data contains mixed occlusion types, blindly reducing the distances between images with different occlusion types could have negative impact. The network may not effectively capture occlusion-specific attention maps. Qualitative results in Figure 4 show that the vanilla triplet loss focuses on areas around the body parts, including the occluding objects. Table 4 also shows clear improvements up to 10% when using grouped triplet loss.

(a) Images

(b) $\mathcal{L}_{\text{tri}}$

(c) $\mathcal{L}_{\text{tri-g}}$

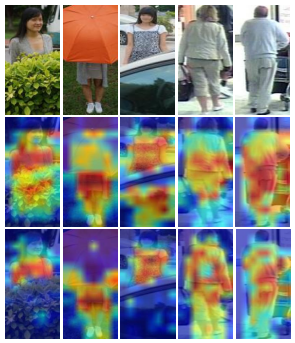

Figure 4: Attention heat maps of different triplet losses.

Type of Triplet	Occluded-Duke		Partial-REID		Partial-iLIDS	
	Rank-1	*m*AP	Rank-1	*m*AP	Rank-1	*m*AP
$\mathcal{L}_{\text{tri}}$	55.2	39.8	76.7	69.9	73.9	76.6
$\mathcal{L}_{\text{tri-g}}$	**62.6**	**46.1**	**86.0**	**77.4**	**77.3**	**79.5**

Table 4: Comparing performance when using $\mathcal{L}_{\text{tri}}$ or $\mathcal{L}_{\text{tri-g}}$.

4.3.3 Choice of Attentive Layers

We incorporate the attention mechanism by appending the proposed attention-guided mask module to a particular intermediate layer of the backbone network. Our objective is to capture more features describing the contour information. Below, we study how the performance varies among different choices of the intermediate layer. Table 5 demonstrates the performance when putting the attention mechanism at different layers of the backbone network *i.e.*, ResNet50.

Shallow layers (layer-1), though preserve rich contour features, are not sufficiently distinguishable. Deeper layers (layer-4), however, mostly describes texture features that cause the attention on occlusion regions. In contrast, the chosen intermediate layer-3 obtains the best results.

Layer	Occluded-Duke		Partial-REID		Partial-iLIDS	
	Rank-1	*m*AP	Rank-1	*m*AP	Rank-1	*m*AP
layer-1	57.5	43.5	80.7	73.2	60.5	65.0
layer-2	56.9	43.2	82.0	75.5	76.5	79.3
layer-3	**62.6**	**46.1**	**86.0**	**77.4**	**77.3**	**79.5**
layer-4	60.0	45.7	76.3	70.5	73.1	76.0

Table 5: Comparing performance when adding the attention network to different layers. Layer-k means appending to the k-th layer in the backbone network (ResNet-50).

4.4. Ablation Study

We examine the effectiveness of each component: occlusion augmentation, group triplet loss, the constraint of the attention-guided mask module, and the occlusion unification. We report the results in Table 6. In particular, we consider the baseline model: $\mathcal{L}_{\text{ide}}$ and $\mathcal{L}_{\text{tri}}$.

Occlusion Augmentation. We observe over 10% accuracy improvements over the baseline model when enabling the occlusion augmentation. This implies that our scheme can produce diverse occlusion data.

OA	$\mathcal{L}_{\text{tri-g}}$	$\mathcal{L}_{\text{mask}}$	OU	Occluded-Duke		Partial-REID		Partial-iLIDS	
				Rank-1	*m*AP	Rank-1	*m*AP	Rank-1	*m*AP
✗	✗	✗	✗	36.9	26.2	56.7	53.0	62.2	66.8
✓	✗	✗	✗	51.2	37.5	76.0	71.2	67.2	70.2
✓	✓	✗	✗	55.5	41.6	78.7	74.5	71.4	74.9
✓	✓	✓	✗	57.7	44.0	82.0	75.1	75.6	78.8
✓	✓	✗	✓	60.3	44.1	82.3	76.3	73.1	76.7
✓	✓	✓	✓	**62.6**	**46.1**	**86.0**	**77.4**	**77.3**	**79.5**

Table 6: Ablation study of occlusion augmentation (OA), grouped triplet ($\mathcal{L}_{\text{tri-g}}$), attention-guided mask module ($\mathcal{L}_{\text{mask}}$), and occlusion unification (OU).

Grouped Triplet Loss. From the second and third rows, $\mathcal{L}_{\text{tri-g}}$ improves 4.3% accuracy over $\mathcal{L}_{\text{tri}}$, suggesting that the grouped triplet loss can effectively guide the learning of occlusion-aware masks.

Mask Module and Occlusion Unification. We study the attention-guided mask module's effectiveness by comparing the last three rows with the third row. Both $\mathcal{L}_{\text{mask}}$ and OU can help capture a more precise attention, and achieve the best results when both of them are enabled.

5. Conclusion

In this paper, we investigate the occlusion challenge in person ReID. We identify the major weakness in previous approaches for holistic, partial, and occluded person ReID. We propose the Occlusion-Aware Mask Network (OAMN) with three innovative components: attention-guided mask module, occlusion augmentation, and occlusion unification. At the training phase, occlusion augmentation produces diverse and labeled occlusion data to guide the attention-guided mask module. At the test phase, occlusion unification mitigates the query image's ambiguity. In summary, OAMN enables existing attention mechanisms to precisely capture body parts regardless of the occlusion. Comprehensive experiments on a variety of person ReID benchmarks demonstrate the superiority of OAMN over stat-of-the-arts.

Acknowledgment

This work is supported by the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No.U1705262, No.62072386, No.62072387, No.62072389, No.62002305, No.61772443, No.61802324 and No.61702136), Guangdong Basic and Applied Basic Research Foundation (No.2019B1515120049) and the Fundamental Research Funds for the Central Universities (No.20720200077, No.20720200090 and No.20720200091).

References

[1] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. In *CVPR*, pages 1320–1329. IEEE Computer Society, 2017. 2

[2] Husheng Dong, Ping Lu, Shan Zhong, Chunping Liu, Yi Ji, and Shengrong Gong. Person re-identification by enhanced local maximal occurrence representation and generalized similarity metric learning. *Neurocomputing*, 307:25–37, 2018. 1

[3] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. Pose-guided visible part matching for occluded person reid. In *CVPR*, pages 11741–11749. IEEE, 2020. 3, 6, 7

[4] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, and Hongsheng Li. FD-GAN: pose-guided feature distilling GAN for robust person re-identification. In *NeurIPS*, pages 1230–1241, 2018. 6

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016. 6

[6] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *CVPR*, pages 7073–7082. IEEE Computer Society, 2018. 2, 3, 6, 7

[7] Lingxiao He, Zhenan Sun, Yuhao Zhu, and Yunbo Wang. Recognizing partial biometric patterns. *CoRR*, abs/1810.07399, 2018. 6, 7

[8] Lingxiao He, Yinggang Wang, Wu Liu, He Zhao, Zhenan Sun, and Jiashi Feng. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In *ICCV*. 2, 3, 7

[9] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017. 4

[10] Houjing Huang, Xiaotang Chen, and Kaiqi Huang. Human parsing based alignment with multi-task learning for occluded person re-identification. In *ICME*, pages 1–6. IEEE, 2020. 1, 3, 7

[11] Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E. Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, pages 1062–1071. IEEE Computer Society, 2018. 7

[12] Martin Köstinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, pages 2288–2295, 2012. 1

[13] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, pages 2285–2294. IEEE Computer Society, 2018. 2

[14] Shengcai Liao and Stan Z. Li. Efficient PSD constrained asymmetric metric learning for person re-identification. In *ICCV*, pages 3685–3693, 2015. 1

[15] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *CVPR*, pages 4099–4108, 2018. 7

[16] Hao Luo, Xing Fan, Chi Zhang, and Wei Jiang. Stnreid : Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification. *CoRR*, abs/1903.07072, 2019. 3

[17] Hao Luo, Wei Jiang, Xing Fan, and Chi Zhang. Stnreid: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification. *IEEE Trans. Multim.*, 22(11):2905–2913, 2020. 2

[18] Bingpeng Ma, Yu Su, and Frédéric Jurie. Covariance descriptor based on bio-inspired features for person re-identification and face verification. *Image Vis. Comput.*, 32(6-7):379–390, 2014. 2

[19] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *ICCV*, pages 542–551, 2019. 2, 3, 6, 7

[20] Lei Qi, Jing Huo, Lei Wang, Yinghuan Shi, and Yang Gao. Maskreid: A mask based deep ranking neural network for person re-identification. *CoRR*, abs/1804.03864, 2018. 7

[21] Tao Ruan, Ting Liu, Zilong Huang, Yunchao Wei, Shikui Wei, and Yao Zhao. Devil in the details: Towards accurate single and multiple human parsing. In *AAAI*, pages 4814–4821. AAAI Press, 2019. 3, 6, 7

[22] Sebastian Ruder. An overview of gradient descent optimization algorithms. *CoRR*, abs/1609.04747, 2016. 6

[23] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C. Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR*, pages 5363–5372, 2018. 7

[24] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, pages 1179–1188, 2018. 7

[25] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, pages 3980–3989, 2017. 7

[26] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *CVPR*, pages 393–402. Computer Vision Foundation / IEEE, 2019. 3, 7

[27] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and A strong convolutional baseline). In *ECCV*, pages 501–518, 2018. 2, 6, 7

[28] Hongchen Tan, Xiuping Liu, Shengjing Tian, Baocai Yin, and Xin Li. Mhsa-net: Multi-head self-attention network for occluded person re-identification. *CoRR*, abs/2008.04015, 2020. 2, 7

[29] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. Aanet: Attribute attention network for person re-identifications. In *CVPR*, pages 7134–7143. Computer Vision Foundation / IEEE, 2019. 2

[30] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *CVPR*, pages 6448–6457, 2020. 2, 3, 6, 7

[31] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. GLAD: global-local-alignment descriptor for pedestrian retrieval. In Qiong Liu, Rainer Lienhart, Haohong Wang, Sheng-Wei "Kuan-Ta" Chen, Susanne Boll, Yi-Ping Phoebe Chen, Gerald Friedland, Jia Li, and Shuicheng Yan, editors, *ACM MM*, pages 420–428. ACM, 2017. 2

[32] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z. Li. Salient color names for person re-identification. In *ECCV*, pages 536–551, 2014. 1, 2

[33] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *CVPR*, pages 3183–3192, 2020. 1, 2, 3, 4, 6

[34] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, pages 907–915. IEEE Computer Society, 2017. 2

[35] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose-invariant embedding for deep person re-identification. *IEEE Trans. Image Process.*, 28(9):4500–4509, 2019. 2

[36] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124. IEEE Computer Society, 2015. 6

[37] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jian-Huang Lai, and Shaogang Gong. Partial person re-identification. In *ICCV*, pages 4678–4686. IEEE Computer Society, 2015. 3, 6

[38] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *ICCV*, pages 3774–3782. IEEE Computer Society, 2017. 6

[39] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, pages 3652–3661. IEEE Computer Society, 2017. 2

[40] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, pages 13001–13008, 2020. 2, 3, 6

[41] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. Occluded person re-identification. In *ICME*, pages 1–6. IEEE Computer Society, 2018. 3

[42] Jiaxuan Zhuo, Jianhuang Lai, and Peijia Chen. A novel teacher-student learning framework for occluded person re-identification. *CoRR*, abs/1907.03253, 2019. 2, 3, 6, 7