Attention Disturbance and Dual-Path Constraint Network for Occluded Person Re-identification

Jiaer Xia[1*], Lei Tan[1*], Pingyang Dai[1†], Mingbo Zhao[2], Yongjian Wu[3], Liujuan Cao[1]

[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, China
[2]Donghua University, Shanghai, China
[3]Tencent Youtu Lab, Shanghai, China
{xiajiaer, tanlei}@stu.xmu.edu.cn, pydai@xmu.edu.cn,
mzhao4@dhu.edu.cn, littlekenwu@tencent.com, caoliujuan@xmu.edu.cn

Abstract

Occluded person re-identification (Re-ID) aims to address the potential occlusion problem when matching occluded or holistic pedestrians from different camera views. Many methods use the background as artificial occlusion and rely on attention networks to exclude noisy interference. However, the significant discrepancy between simple background occlusion and realistic occlusion can negatively impact the generalization of the network. To address this issue, we propose a novel transformer-based Attention Disturbance and Dual-Path Constraint Network (ADP) to enhance the generalization of attention networks. Firstly, to imitate real-world obstacles, we introduce an Attention Disturbance Mask (ADM) module that generates an offensive noise, which can distract attention like a realistic occluder, as a more complex form of occlusion. Secondly, to fully exploit these complex occluded images, we develop a Dual-Path Constraint Module (DPC) that can obtain preferable supervision information from holistic images through dual-path interaction. With our proposed method, the network can effectively circumvent a wide variety of occlusions using the basic ViT baseline. Comprehensive experimental evaluations conducted on person re-ID benchmarks demonstrate the superiority of ADP over state-of-the-art methods.

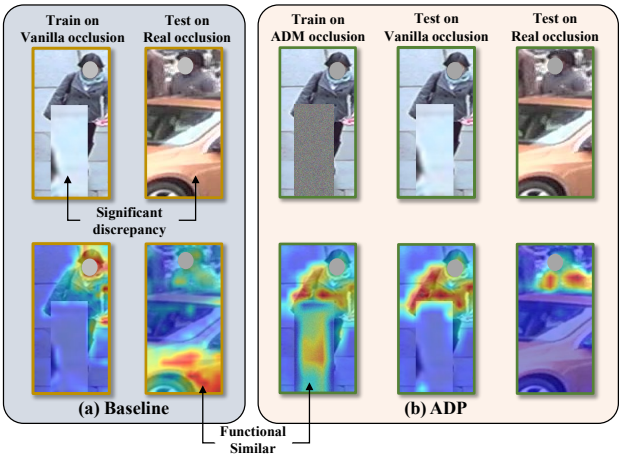


Figure 1: Visualization of attention to baseline and proposed ADP. (a) The baseline trained with the assistance of background occlusion failed to avoid the realistic occlusion in the testing set. (b) The ADP trained by the proposed reality-similar occlusion ADM performs well on both artificial and real occlusion.

Introduction

Person re-identification (Re-ID) refers to the process of matching pedestrian images captured by non-overlapping cameras. This technique has gained popularity in recent years as surveillance systems have become more advanced and widespread. With the rapid development of deep learning technology (He et al. 2016; Vaswani et al. 2017; Dosovitskiy et al. 2021), Re-ID has also achieved remarkable performance (Luo et al. 2019; Zhai et al. 2020; Zheng et al. 2015; Eom and Ham 2019; Chen et al. 2018; Wu et al. 2016) by meriting from its powerful feature extraction capabilities. However, most existing methods assume that the pedestrians in retrieved images are unobstructed, ignoring the possible occlusion problems that can occur in real-world scenarios. Consequently, these methods significantly degrade when dealing with occluded images. While recent endeavors have facilitated person Re-ID under occlusion conditions (Yan et al. 2021; Wang et al. 2022b; Tan et al. 2022; Li et al. 2021; Shi et al. 2022; Jia et al. 2022), two main problems associated with occlusions still need to be addressed. Firstly, the presence of obstacles will vanish some parts of the human body, missing and misaligned extracted features. Traditional Re-ID methods cannot perform valid retrievals when some discriminative parts are obscured. Secondly, occlusions introduce noise into extracted features, polluting the final feature representation of each image. When dealing with these polluted features, different identities may have high similarities due to the same obstacle, resulting in incorrect matches. To address the aforementioned problems, some methods (Wang et al. 2020; Gao et al. 2020; Miao et al. 2019; Miao, Wu, and Yang 2022) use additional trained networks, such as human parsing and keypoint estimation,

*Equal contribution.
†Corresponding author.

to align different human parts. With the aid of these extra networks, the occluded parts can be repaired by disseminating information from the visible parts. However, these approaches are severely limited due to the domain gap between the pre-trained network and the Re-ID dataset.

Recently, with the exploration of attention mechanisms for various vision tasks, it has also been adopted for occluded person Re-ID to eliminate the interference of noisy information (Zhao et al. 2021; Sun et al. 2019; He et al. 2021). During the process of attention learning, many data augmentation strategies (Chen et al. 2021; Wang et al. 2022b; Zhuo et al. 2018) generate artificial occlusion, which directs the attention to person and forces it to avoid occluded regions. Currently, the most widely used artificial occlusion methods are random erasing (Zhong et al. 2020) or using the background as occlusion (Chen et al. 2021). Nevertheless, pre-trained attention networks are inherently more likely to focus on the semantically rich foreground than the background. Therefore, network will inevitably tend to ignore the occlusion constituted by the background, which will result in a lack of generalization. To illustrate this point, we utilized background for artificial occlusion based on the ViT baseline in TransReID (He et al. 2021) and visualize the attention for both the training and testing sets in Fig.1(a). The results demonstrate that while the baseline can avoid artificial occlusions well in the training set, attention is still disturbed in the testing set due to the significant discrepancy between artificial occlusion and actual occlusion.

In this paper, we propose a solution to the challenges mentioned above by introducing an Attention Disturbance Mask (ADM) module that simulates real-world occlusions with greater fidelity. The primary way in which occlusions disrupt models is by impeding attention. However, obtaining enough occluded data to enable the model to avoid such disruptions is difficult. To surmount this problem, we utilize an attack-oriented methodology that produces noise masks with the capacity to simulate the interference effects of actual obstructions at the feature level. This enables us to construct occlusions that mirror the effects of those encountered in real-world scenarios. As illustrated in Figure 1(b), the proposed Attention Disturbance Module (ADM) performs a similar role to real-world occlusions by introducing disruptions to the neural network's attention. This finding directly verifies the capability of our designed ADM in faithfully emulating occlusions at the feature level. By training the network on such occlusions that closely resemble those encountered in real-world scenarios, we can effectively enhance its robustness against occlusions during testing.

However, handling complex occlusions directly can pose optimization challenges for the network. To address this issue, we propose the Dual-Path Constraint Module (DPC) to handle both holistic and occluded images simultaneously, thus using holistic features as an extra supervisor to guide attention more towards the target pedestrian. Notably, the network parameters in the proposed DPC are shared by both paths, while the individual classifiers learn information about holistic and occluded images separately.

The main contributions of our method can be summarized as below:

- We first introduce a novel attack-based augmentation strategy called the Attention Disturbance Mask (ADM), which simulates real occlusion at the feature level and effectively diverts attention away from actual occlusions during testing.
- We propose a Dual-Path Constraint module (DPC) that utilizes dual-path interactions to encourage the network to learn a more generalized attention mechanism. DPC is compatible with existing occlusion-based data augmentation methods and can provide significant performance improvements.
- The two proposed methods are both used to assist in the training of the baseline, and can be discarded in the inference stage, making them easy to be compatible with many existing methods, indicating the efficiency and wide applicability of our method.
- Trained with our proposed ADP, the transformer baseline can achieve new state-of-the-art performance on multiple benchmark datasets *e.g.*, 74.5% on Rank-1 on Occluded-Duke dataset.

Related Work

Two main issues of occluded person Re-ID are the missing information and the noisy information caused by the various obstacles. Some methods have been proposed to address the missing information issue by attempting to remove the obstacle and reconstruct the missing visible human parts (Hou et al. 2019, 2022; Wang et al. 2020). Hou *et al.* (Hou et al. 2019) designed an auto-encoder to generate contents of the occluded parts at pixel level. Wang *et al.* (Wang et al. 2020) utilize an additional pose estimate network to detect the keypoints then find the high-order relation and human-topology information. Moreover, an adaptive direction graph convolutional layer is proposed to pass relation information from visible to occluded nodes. Hou *et al.* (Hou et al. 2022) divide the feature map into six regions and predict the features of occluded regions by adopting the long-range spatial contexts from non-occluded regions. Instead of reconstructing the missing parts, other methods choose to more focus on the visible parts. With the help of the attention mechanism, these methods can generate attention maps in which occluded regions are given smaller weights to discard the noisy information. To better discover the occluded part, several data augmentation strategies are adopted. Zhuo *et al.* (Zhuo et al. 2018) design an occlusion simulator to use the random patch from the background as the artificial occlusion to cover the full-body person image. Chen *et al.* (Chen et al. 2021) also use the background information and paste it to eight predefined locations to make the data augmentation.

These artificial occlusion data augmentation can provide labels for the location of occlusions. This extra information helps train attention mechanisms to exclude noisy occlusions. However, networks trained on artificial occlusions often cannot handle realistic occlusions well, due to the significant discrepancy between artificial and realistic occlusion. The proposed ADP method takes a different approach. It generates more realistic occlusions while also providing

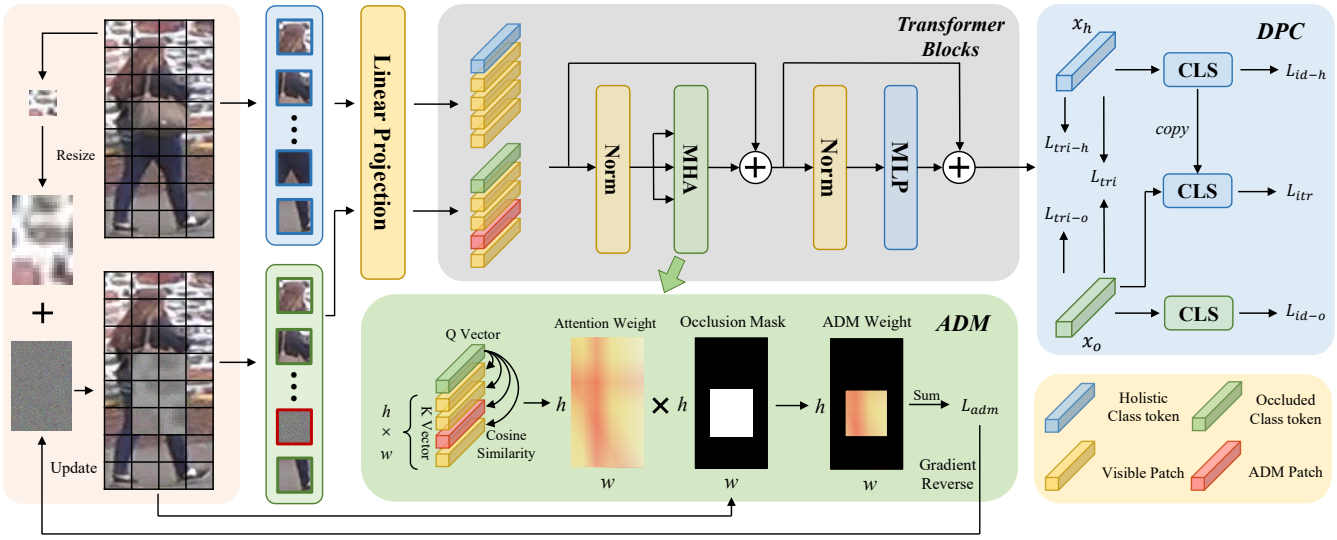


Figure 2: The overview of the proposed Attention Disturbance and Dual-Path Constraint Network (ADP). To create the corresponding occlusion images, the transformed background patch is used as the carrier of the Attention Disturbance Mask (ADM) and covers a random region of the original image. Then the Dual-Path Constraint Module (DPC) simultaneously deals with holistic and occluded images. In Multi-Head Attention (MHA) stage, ADM maximizes the similarity between class token and masked patches to optimize the mask.

extra supervision from holistic features. The more realistic occlusions lead to more robust and generalized networks.

Proposed Method

Overview

In this section, we introduce the proposed Attention Disturbance and Dual-Path Constraint Network (ADP). The architecture of ADP is depicted in Fig.2, where a pre-trained ViT (Dosovitskiy et al. 2021) is utilized as the backbone to extract image features. To generate occluded images, we use an artificial occlusion consisting of a background patch and an attention disturbance mask. Both holistic and occluded images are fed into a parameter-shared transformer to extract their respective features. Similar to TransReID (He et al. 2021), we divide the input images, which have a resolution of H × W, into N patches and then convert them into patch embeddings via a linear projection operation with a patch size of P and a stride size of S.

$$N = h \times w = \left\lfloor \frac{H+S-P}{S} \right\rfloor \times \left\lfloor \frac{W+S-P}{S} \right\rfloor. \quad (1)$$

Meanwhile, a learnable class embedding token denoted as x_{cls} is attached to the input sequences to aggregate the image information during the process and act as the global feature map in the output stage. Besides, a learnable position embedding $\mathcal{P} \in \mathbb{R}^{(N+1)\times d}$ is added to the transformer to append spatial information to the transformer. The complete input sequence embeddings can be formulated as:

$$\mathcal{Z}_0 = \left[x_{\text{cls}}; \mathcal{F}\left(x_p^1\right); \mathcal{F}\left(x_p^2\right); \cdots; \mathcal{F}\left(x_p^N\right)\right] + \mathcal{P}, \quad (2)$$

where $\mathcal{Z}_0$ represents input sequence embeddings, $\mathcal{F}$ is the linear projection operation mapping the input image $x \in \mathbb{R}^{H\times W\times C}$ into patch embedding $f_{pe} \in \mathbb{R}^{N\times d}$, N is the number of flattened $h \times w$ patched, and d denotes the dimensions. Then, the patch embedding f_{pe} is combined with the class token as the input feature map of transformer blocks.

Attention Disturbance Mask

The main idea of ADM is to generate a mask that can make the network's attention accidentally focus on the occlusion, which simulates the same effect as real-world occlusion. However, directly generating a mask to make attention focus on a blank area is difficult. Therefore, the background information is adopted as a carrier of mask to provide the domain information and simplify the mask optimization process. To get the background region of the input image, we randomly select the corner patch of image. Then the cropped background patch P_b will be resized to $s_o = r_o \times s$, where $r_o \sim \mathcal{U}(0.1, 0.5)$ and $s = H \times W$. The shape of P_b is $H_b = \sqrt{s_o \times r_s}$ and $W_b = \sqrt{\frac{s_o}{r_s}}$ with $r_s \sim \mathcal{U}(0.3, 3.3)$. The processed background patch is pasted arbitrarily anywhere in the image, and a mask $\mathcal{M} \in \mathbb{R}^{H\times W}$ corresponding to the pasting position is saved for overlay attention disturbance mask.

To generate the ADM, we first initialize a random learnable parameter, which has same shape as input image, and superimpose it to the occlusion position according to the $\mathcal{M}$. During training, we dynamically update the ADM based on the weight matrix in the multi-head attention stage of each transformer block. Considering the mechanism of attention operation, it first calculates the dot-product between the queries Q and the keys K to measure the similarity between them and in accordance with the similarity to get the

weight matrix (Vaswani et al. 2017). The whole attention process can be formulated as follows:

$$\text{Attention}(Q, K, V) = \mathbf{W}V, \tag{3}$$

$$\mathbf{W} = softmax(\frac{\mathbf{QK^T}}{\sqrt{c/N_H}}), \tag{4}$$

where N_H means the number of heads in multi-head attention.

Therefore, we can disturb the attention by maximizing the similarity between the class token and the occluded region, which will increase the occluded region's weight and make attention mistakenly focus on it. To be specific, in each transformer block, we have class token embedding x^i_{cls} and patch embedding x^i_p, where i represents the i-th block of the transformer. Then the disturbance loss can be formulated as:

$$\mathcal{L}_{adm} = \sum_{i=1}^{N} \sum_{n} softmax(\frac{\mathbf{x^i_{cls}} \mathbf{x^i_p}^{\mathbf{T}}}{\sqrt{c/N_H}}) \odot \tilde{\mathcal{M}}, \tag{5}$$

$\tilde{\mathcal{M}}$ represents the resized occlusion mask $\mathcal{M}$ according to the patch size, $\odot$ presents element-wise product. Then we adopt an additional optimizer to update the ADM alone based on the reversed gradient of the disturbance loss.

Dual-Path Constraint Module

By generating a corresponding occluded image for each input image, we can obtain the holistic-occluded paired data and make it possible to exploit the holistic image as additional supervision for occluded images, which will help deal with complex occlusion cases. We separately use the identity loss and metric loss in both holistic and occluded paths to ensure the extracted features towards respective images are reliable and discriminative. Meanwhile, a global metric loss and information passing classifier is adopted to convey information between two paths.

After extracting the features of holistic and occlusion images, we obtain the final class token of each path as the feature map denoted as x_h and x_o.

In the holistic path, we use the cross-entropy loss as ID-loss and triplet loss as metric loss. The id-loss $\mathcal{L}_{id-h}$ and metric loss $\mathcal{L}_{tri-h}$ are shown as follows:

$$\mathcal{L}_{id-h} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{e^{(W_h^{y_i})^T x_h^i}}{\sum_{j=1}^{C} e^{(W_h^{y_j})^T x_h^i}}, \tag{6}$$

$$\mathcal{L}_{tri-h} = \left[\alpha + \left(\|f_a - f_p\|_2^2 - \|f_a - f_n\|_2^2\right)\right]_+, \tag{7}$$

where the B and C in ID-loss refer to the batch size and the number of class, and W_h represents the weight of holistic classifier. The f_a, f_p, and f_n in triplet loss refer to the anchor, positive and negative features with online hard-mining (Schroff, Kalenichenko, and Philbin 2015), and α is the margin. The loss of holistic path can be calculated as:

$$\mathcal{L}_h = \mathcal{L}_{id-h} + \mathcal{L}_{tri-h}. \tag{8}$$

In the occluded path, the existence of occlusion weakens the identity information and fuzzes the inter-class discrepancy when they have the same occlusion. The widely used softmax loss is incapable of achieving splendid enough intra-class compactness in such difficult conditions. So we adopt an extra angular margin in the original softmax loss to increase intra-class compactness and inter-class discrepancy (Deng et al. 2019; Tan et al. 2022). The id-loss $\mathcal{L}_{id-o}$ with the extra margin can be represented as:

$$\mathcal{L}_{id-o} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{e^{s(\theta_i + m)}}{e^{s(\theta_i+m)} + \sum_{j=1, j\neq y_i}^{C} e^{s(\theta_j)}}, \tag{9}$$
$$\theta_i = (W_o^{y_i})^T x_o^i, \;\; \theta_j = (W_o^{y_j})^T x_o^j,$$

where m denotes the angular margin and s is the scale adjust hyper-parameter. And metric loss $\mathcal{L}_{tri-o}$ is also adopted, which is same as holistic path. The loss of occluded path can be calculated as:

$$\mathcal{L}_o = \mathcal{L}_{id-o} + \mathcal{L}_{tri-o}. \tag{10}$$

To connect the holistic and occluded path, we adopt a global triplet loss $\mathcal{L}_{tri}$ to close the distance between these two paths. Furthermore, since the classifier can be regarded as a prototype center of each identity, we use it as an anchor of the holistic identity feature to pull close the same identity in the occluded path to mitigate the gap between them. And benefits from the asymmetric structure of the classifier in two paths, the interaction between two paths will bring more information and provide more substantial supervision. Specifically, we clone the parameter of the classifier in holistic path as $\hat{W}_h$ and calculate the similarity with occluded feature as the interaction loss to eliminate the effect of occlusion. The interaction loss $\mathcal{L}_{itr}$ can be given as:

$$\mathcal{L}_{itr} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{e^{(\hat{W}_h^{y_i})^T x_o^i}}{\sum_{j=1}^{C} e^{(\hat{W}_h^{y_j})^T x_o^i}}. \tag{11}$$

The whole loss function of Dual-Path Constraint Module can be summarized as:

$$\mathcal{L}_{dpc} = \mathcal{L}_h + \mathcal{L}_o + \mathcal{L}_{tri} + \lambda \mathcal{L}_{itr}, \tag{12}$$

where the λ is the hyper-parameter coefficient of $\mathcal{L}_{itr}$.

In the testing phase, only the pure ViT baseline is used to extract feature maps without any artificial occlusion, which makes our network simple and efficient for implementation.

Experiment

Datasets and Evaluation Setting

To validate the effectiveness of our proposed method, we perform extensive experiments on publicly available Re-ID datasets, including both occluded (Miao et al. 2019; Zhuo et al. 2018) and holistic (Zheng et al. 2015; Zheng, Zheng, and Yang 2017; Ristani et al. 2016) datasets.

Occluded-Duke (Miao et al. 2019) is a large-scale dataset selected from the DukeMTMC for occluded person re-identification. It consists of 15,618 training images of 702 people, while the query and gallery sets contain 2,210 testing images of 519 people and 17,661 images of 1,110 persons, respectively. Until now, Occluded-Duke is still the

Methods	Occ-Duke		Occ-REID	
	R-1	*m*AP	R-1	*m*AP
PCB(Sun et al. 2018)	42.6	33.7	41.3	38.9
DSR(He et al. 2018)	40.8	30.4	72.8	62.8
FPR(He et al. 2019)	-	-	78.3	68.0
Ad-Occ(Huang et al. 2018)	44.5	32.2	-	-
PVPM(Gao et al. 2020)	47.0	37.7	66.8	59.5
GASM(He and Liu 2020)	-	-	74.5	65.6
HOReID(Wang et al. 2020)	55.1	43.8	80.3	70.2
OAMN(Chen et al. 2021)	62.6	46.1	-	-
Part-Label(Yang et al. 2021)	62.2	46.3	81.0	71.0
ISP(Zhu et al. 2020)	62.8	52.3	-	-
PRE-Net(Yan et al. 2023)	68.3	55.2	-	-
CAAO(Zhao et al. 2023)	68.5	59.5	87.1	83.4
PAT(Li et al. 2021)	64.5	53.6	81.6	72.1
TransReID(He et al. 2021)	64.2	55.7	-	-
PFD(Wang et al. 2022a)	67.7	60.1	79.8	81.3
FED(Wang et al. 2022b)	68.1	56.4	86.3	79.3
SAP(Jia et al. 2023)	70.0	62.2	83.0	76.8
ADP(Ours)	72.2	60.6	88.2	82.0
TransReID*(He et al. 2021)	66.4	59.2	-	-
PFD*(Wang et al. 2022a)	69.5	61.8	81.5	83.0
DPM*(Tan et al. 2022)	71.4	61.8	85.5	79.7
ADP(Ours)*	**74.5**	**63.8**	**89.2**	**85.1**

Table 1: Comparison with state-of-the-art methods on Occluded-Duke and Occluded-REID. $*$ indicates that the backbone has a sliding-window setting and a smaller stride.

Methods	Market-1501		DukeMTMC	
	R-1	*m*AP	R-1	*m*AP
PCB(Sun et al. 2018)	92.3	77.4	81.8	66.1
ISP(Zhu et al. 2020)	95.3	88.6	89.6	80.0
BOT(Luo et al. 2019)	94.1	85.7	86.4	76.4
DSR(He et al. 2018)	50.7	70.0	58.8	67.2
STNReID(Luo et al. 2020)	66.7	80.3	54.6	71.3
VPM(Sun et al. 2019)	93.0	80.8	83.6	72.6
HOReID(Wang et al. 2020)	94.2	84.9	86.9	75.6
OAMN(Chen et al. 2021)	93.2	79.8	86.3	72.6
FPR(He et al. 2019)	95.4	86.6	88.6	78.4
PAT(Li et al. 2021)	95.4	88.0	88.8	78.2
FED(Wang et al. 2022b)	95.0	86.3	89.4	78.0
TransReID$*$(He et al. 2021)	95.2	88.9	90.7	82.0
PRE-Net(Yan et al. 2023)	95.3	86.5	89.3	77.8
CAAO(Zhao et al. 2023)	95.3	88.0	89.8	80.9
DPM$*$(Tan et al. 2022)	95.5	89.7	91.0	82.6
PFD$*$(Wang et al. 2022a)	95.5	89.7	**91.2**	**83.2**
SAP(Jia et al. 2023)	**96.0**	**90.5**	-	-
ADP(Ours)$*$	95.6	89.5	**91.2**	83.1

Table 2: Comparison with state-of-the-art methods on Market-1501 and DukeMTMC-reID. $*$ indicates that the backbone has a sliding-window setting and a smaller stride.

most challenging dataset for occluded Re-ID due to the scale of occlusion.

Occluded-REID (Zhuo et al. 2018) is an occluded person dataset captured by mobile cameras. A total of 2,000 images were captured from 200 individuals, each consisting of five full-body images and five occluded images. Following the evaluation protocol of previous works (Gao et al. 2020; Wang et al. 2020), we trained the model under the training set of Market-1501 (Zheng et al. 2015),while Occluded-REID is used only as a test set.

Market-1501 (Zheng et al. 2015) is a widely-used holistic Re-ID dataset captured from 6 cameras. The training set contains 1,236 images of 751 people, while the query and gallery sets contain 3,368 images of 750 people and 19,732 images of 750 people, respectively.

DukeMTMC-reID (Zheng, Zheng, and Yang 2017; Ristani et al. 2016) contains 36,441 images of 1,812 persons captured by 8 cameras, with 16,522 images of 702 identities are used as the training set and 2,228 and 16,522 images of 702 people who do not appear in the training set as the query and gallery images, respectively.

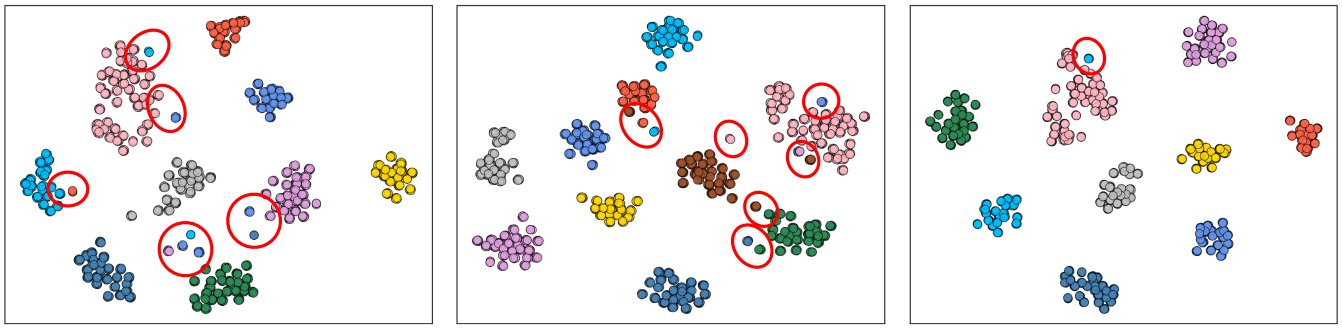
Evaluation Protocol. To make it fair compared with other methods, we adopt the widely-used Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) as evaluation metrics. All experiments are performed in the single query setting.

Implementation details. We adopt the ViT (Dosovitskiy et al. 2021) pre-trained on ImageNet (Deng et al. 2009) as our backbone and use 12 transformer blocks with 8 heads for multi-head attention. The numbers of channel is set to 768. The input images are resized to 256×128 and augmented by commonly used random horizontal flipping, padding and random cropping. During the training phase, the batch size is set to 64 with 16 identities. We utilize the SGD as the optimizer, with the initial learning rate of 0.004 and a cosine learning rate decay. The margin of each triplet loss is set to 0.3. The hyper-parameter m and s in eq.(9) are set to 0.3 and 30, respectively, while the λ in eq.(12) is 0.1.

Comparison with State-of-the-art Methods

Result on Occluded Datasets. We compare our ADP with existing state-of-the-art (SOTA) methods on two occluded datasets, and the results are shown in Table 1. The comparison methods can be divided into CNN-based methods and Transformer-based methods. As can be seen from Table 1, transformer-based methods outperform the CNN-based methods. This improvement can be achieved by approximately 15%, which demonstrates the utilization of attention mechanism is beneficial to occlusion tasks. A case in point is that in the most challenging Occluded-Duke dataset, our proposed method ADP can achieve 72.2% in rank-1. Furthermore, with a small step sliding-window setting, the proposed ADP* can further achieve a higher performance of 74.5% in rank-1 and 63.8% in *m*AP, respectively, exceeds +3.1% in Rank-1 and +2.0% in *m*AP compared with the transformer-based SOTA method DPM (Tan et al. 2022). On the Occluded-REID dataset, our ADP and ADP* also consistently outperform current SOTAs.

Result on Holistic Datasets. We also experiment our proposed method on holistic person Re-ID datasets, including Market-1501 and DukeMTMC-reID, and compare our method with state-of-the-art methods in three categories, *i.e.*, holistic Re-ID methods (Zhu et al. 2020; Luo et al. 2019; Sun et al. 2018), partial Re-ID methods (Luo et al.



(a)Baseline **(b) Occluded by background** **(c) Occluded by ADM**

Figure 3: Visualization of the feature distribution on Occlude-Duke dataset. Circles denote the features of images while the colors represent different identities. (a) Baseline refers to the model trained on the images without extra occlusions. (b) The middle plot shows the results of model trained on the images occluded by the background. (c) Compared to the other two models, the model trained on our ADM can avoid the influence of obstacles well.

Method	Occluded-Duke			
	R-1	R-5	R-10	*m*AP
baseline	59.7	75.3	80.7	49.8
+ADM	66.2	81.6	86.3	57.7
+DPC	72.2	85.1	88.0	60.6
baseline*	63.2	78.8	83.6	53.3
+ADM	69.8	82.8	87.5	60.3
+DPC	**74.5**	**86.4**	**89.6**	**63.8**

Table 3: Ablation study of each proposed module in ADP on Occluded-Duke dataset. $*$ indicates that the backbone has a sliding-window setting and a smaller stride.

Method	Occluded-Duke			
	R-1	R-5	R-10	*m*AP
baseline	59.7	75.3	80.7	49.8
+ADM	66.2	81.6	86.3	57.7
+AM	69.5	82.2	85.9	58.8
+DP	70.5	83.4	87.0	59.3
+L_{itr}	71.8	83.3	87.1	59.7
+L_{tri}(full)	**72.2**	**85.1**	**88.0**	**60.6**

Table 4: Ablation study of the dual-path loss used in DPC on Occluded-Duke dataset. AM denotes the module use shared angular softmax with single-path structure, while DP represents the module with an asymmetric dual-path structure.

2020; Sun et al. 2019) and occluded Re-ID methods (Wang et al. 2020; Chen et al. 2021; He et al. 2019; Li et al. 2021; Wang et al. 2022b; He et al. 2021; Tan et al. 2022; Wang et al. 2022a; Jia et al. 2023; Yan et al. 2023; Zhao et al. 2023) methods. The results are shown in Table 2. Though designed for occlusion problems, our proposed module achieves comparable performance on holistic datasets. For example, the proposed ADP can achieve +0.3%/+1.6% improvement in Rank-1 and +0.9%/+3.1% in *m*AP on Market-1501 and DukeMTMC-ReID datasets, respectively, compared with the state-of-the-art method ISP (Zhu et al. 2020). Our ADP also got +2.6%/+7.6% improvement in Rank-1 and +8.7%/+10.5% in *m*AP compared with SOTA partial Re-ID method VPM (Sun et al. 2019).

Ablation Studies

In this section, we implement the ablation studies based on the Occluded-Duke dataset to analyze the influence of each module of the proposed ADP method.

In our study, the baseline method adopts ViT as the backbone of network, which is trained based on the original softmax loss and triplet loss without any artificial occlusion. The results of ablation studies are given in Table 3. From the result, we can observe that training with images occluded by the ADM can significantly improve the model performance, in a way that the performance can be increased by +6.5% in Rank-1 and +7.9% in *m*AP, respectively, over baseline. Meanwhile, this improvement of performance can reach +6.6% in Rank-1 and +7.0% in *m*AP, respectively, over baseline* which only has a smaller stride. Besides, with the assistance of DPC, the performance of the model can further increase from 66.2% to 72.2% in Rank-1 and 57.7% to 60.6% in *m*AP over baseline, and increase from 69.8% to 74.5% in Rank-1 and 60.3% to 63.8% in *m*AP over baseline*.

We next conduct the ablation test to evaluate the influence of structure and loss function in the DPC module on Occluded-Duke dataset. The result is given in Table 4, which shows the effectiveness of the proposed dual-path structure and adopted loss used for connecting two paths. Specifically, with the single-path structure, the adopted angular margin softmax does improve the performance, but it is not suitable for dealing with holistic images and is prone to overfitting problem. To the contrast, since we propose the dual-path structure and separate the holistic and occluded images to use an asymmetric classification, the performance increased by +1.0% in Rank-1 and +0.5% in *m*AP. Meanwhile, bene-

Index	RE	BG	ADM	DPC	Rank-1	*m*AP
1					59.7	49.8
2	✓				61.1	53.8
3		✓			64.8	56.5
4			✓		**66.2**	**57.7**
5	✓			✓	66.7	56.5
6		✓		✓	70.4	58.4
7			✓	✓	**72.2**	**60.6**

Table 5: Comparison with previous occlusion strategies. RE indicates the random erasing method, while BG denotes the occlusion strategy using background.

fiting from the proposed dual-path structure, we can add additional connections between two paths to better leverage the advantages of different types of data. As a result, L_{itr} and L_{tri} can improve the performance of rank1 by +1.3%/+0.4% and *m*AP by +0.4%/+0.9%, respectively.

Discussions

Effectiveness of ADM occlusion. To better demonstrate the advantages of ADM occlusion, we compare various occlusion schemes, including random erasing (Zhong et al. 2020) and directly using the background as occlusion (Chen et al. 2021). The results are shown in Table 5. From index-2 to index-4, ADM exhibits prominent performance improvement over conventional occlusion schemes. In detail, compared with random erasing, the performance is significantly increased by +5.1% in Rank-1 and +3.9% in *m*AP, respectively; compared with the background occlusion, ADM can also achieve +1.4% performance increase in Rank-1 and +1.2% in *m*AP, respectively. Meanwhile, we further visualize the distributions of different features extracted by the model trained with different occlusion strategies. The simulation results are shown in Fig.3, where the circles denote the features of images randomly selected from testing set of Occluded-Duke dataset and visualized via t-SNE (Van der Maaten and Hinton 2008). In detail, Fig.3(a) illustrates the distribution of features extracted by the baseline. It is evident that there are numerous outlier features caused by occlusions. In Fig.3(b), the widely used background occlusion is able to reduce the situation of the outlier features to some extent, but it still cannot completely eliminate them, indicating the model is still affected by the obstacles; To the contrast, with the model trained by our proposed ADM, the outlier features in Fig.3(c) almost disappear due to the model's excellent ability to avoid obstacles. In a nutshell, Fig.3 proves that ADM can help the model reduce the impact of obstacles.

Effectiveness of DPC module. We evaluated the effectiveness of our proposed DPC with different occlusion strategies, and the simulation results are presented in index-5 to index-7 of Table 5. Compared with the result in index-2 to index-4, our experimental results demonstrate that the DPC can be seamlessly integrated into each occlusion strategy and show improvements in performance. For example, in the random erasing strategy, performance can be increased from 61.1% to 66.7% in Rank-1 and 53.8% to 56.5% in *m*AP, respectively. Moreover, with the background occlusion, the DPC can significantly increase the Rank-1 performance by +5.6% and *m*AP performance by +1.9%, respectively. Our results indicate that the proposed DPC has strong compatibility and universality, making it capable of improving the performance of several previous methods.

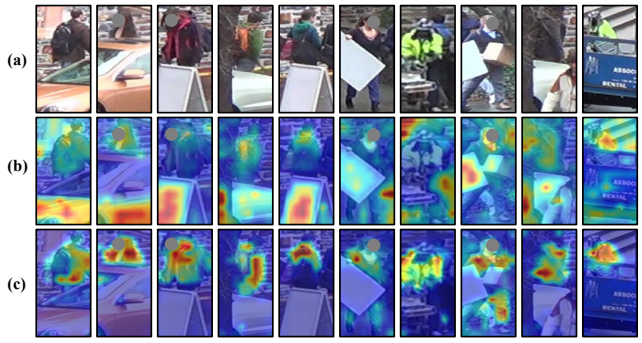


Figure 4: Visualization of attention maps on occlusion testing set in Occluded-Duke dataset. (a) occluded person images. (b) attention maps of baseline. (c) attention maps of ADP.

Visualization of the Attention

To demonstrate the model's ability to process images with occlusions, we visualize the attention maps and show in Fig.4. The input images are from the testing set with diverse occlusions, and we apply Grad-CAM (Selvaraju et al. 2017) to visualize the attention heatmap to demonstrate the areas the model focuses on. It is obvious that baseline can be easily interfered by obstacles, which greatly limits the performance. In the contrast, Fig.4(c) appears that our model's attention mechanism is capable of avoiding paying attention to occlusions to a great extent and focusing more on the target pedestrian. Furthermore, the attention heatmap shows the proposed model can provide good performance when handling diverse occlusion types and locations.

Conclusion

In this research, we introduced a new approach to address the problem of occluded person re-identification by proposing two innovative modules. The ADM generates a more effective artificial occlusion that closely resembles real-world occlusions at the feature level, making the network robust to unseen occlusions and enhancing its generalization. The DPC handles both holistic and occluded images simultaneously, aligning the holistic and occluded features and guiding attention more toward the target pedestrian. Meanwhile, the two proposed modules, ADM and DPC, can be seamlessly integrated with various existing models to enhance their performance, demonstrating the wide applicability of our approach. Experiment results on two occluded datasets and two holistic datasets, illustrate the effectiveness of proposed method and superiority to other state-of-the-art methods.

Acknowledgements

This work was supported by National Key R&D Program of China (No.2022ZD0118202), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. U22B2051, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, No. 62002305 and No. 62272401), and the Natural Science Foundation of Fujian Province of China (No.2021J01002, No.2022J06001).

References

Chen, P.; Liu, W.; Dai, P.; Liu, J.; Ye, Q.; Xu, M.; Chen, Q.; and Ji, R. 2021. Occlude Them All: Occlusion-Aware Attention Network for Occluded Person Re-ID. In *ICCV*, 11813–11822. IEEE.

Chen, Y.; Zhu, X.; Zheng, W.; and Lai, J. 2018. Person Re-Identification by Camera Correlation Aware FFeature Augmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(2): 392–408.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*, 248–255. IEEE Computer Society.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *CVPR*, 4690–4699. Computer Vision Foundation / IEEE.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*. OpenReview.net.

Eom, C.; and Ham, B. 2019. Learning Disentangled Representation for Robust Person Re-identification. In *NeurIPS*, 5298–5309.

Gao, S.; Wang, J.; Lu, H.; and Liu, Z. 2020. Pose-Guided Visible Part Matching for Occluded Person ReID. In *CVPR*, 11741–11749. Computer Vision Foundation / IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778. IEEE Computer Society.

He, L.; Liang, J.; Li, H.; and Sun, Z. 2018. Deep Spatial Feature Reconstruction for Partial Person Re-Identification: Alignment-Free Approach. In *CVPR*, 7073–7082. Computer Vision Foundation / IEEE Computer Society.

He, L.; and Liu, W. 2020. Guided Saliency Feature Learning for Person Re-identification in Crowded Scenes. In *ECCV (28)*, volume 12373 of *Lecture Notes in Computer Science*, 357–373. Springer.

He, L.; Wang, Y.; Liu, W.; Zhao, H.; Sun, Z.; and Feng, J. 2019. Foreground-Aware Pyramid Reconstruction for Alignment-Free Occluded Person Re-Identification. In *ICCV*, 8449–8458. IEEE.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021. TransReID: Transformer-based Object Re-Identification. In *ICCV*, 14993–15002. IEEE.

Hou, R.; Ma, B.; Chang, H.; Gu, X.; Shan, S.; and Chen, X. 2019. VRSTC: Occlusion-Free Video Person Re-Identification. In *CVPR*, 7183–7192. Computer Vision Foundation / IEEE.

Hou, R.; Ma, B.; Chang, H.; Gu, X.; Shan, S.; and Chen, X. 2022. Feature Completion for Occluded Person Re-Identification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(9): 4894–4912.

Huang, H.; Li, D.; Zhang, Z.; Chen, X.; and Huang, K. 2018. Adversarially Occluded Samples for Person Re-Identification. In *CVPR*, 5098–5107. Computer Vision Foundation / IEEE Computer Society.

Jia, M.; Cheng, X.; Lu, S.; and Zhang, J. 2022. Learning disentangled representation implicitly via transformer for occluded person re-identification. *IEEE Transactions on Multimedia*.

Jia, M.; Sun, Y.; Zhai, Y.; Cheng, X.; Yang, Y.; and Li, Y. 2023. Semi-attention Partition for Occluded Person Re-identification. In Williams, B.; Chen, Y.; and Neville, J., eds., *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, 998–1006. AAAI Press.

Li, Y.; He, J.; Zhang, T.; Liu, X.; Zhang, Y.; and Wu, F. 2021. Diverse Part Discovery: Occluded Person Re-Identification With Part-Aware Transformer. In *CVPR*, 2898–2907. Computer Vision Foundation / IEEE.

Luo, H.; Gu, Y.; Liao, X.; Lai, S.; and Jiang, W. 2019. Bag of Tricks and a Strong Baseline for Deep Person Re-Identification. In *CVPR Workshops*, 1487–1495. Computer Vision Foundation / IEEE.

Luo, H.; Jiang, W.; Fan, X.; and Zhang, C. 2020. STNReID: Deep Convolutional Networks With Pairwise Spatial Transformer Networks for Partial Person Re-Identification. *IEEE Trans. Multim.*, 22(11): 2905–2913.

Miao, J.; Wu, Y.; Liu, P.; Ding, Y.; and Yang, Y. 2019. Pose-Guided Feature Alignment for Occluded Person Re-Identification. In *ICCV*, 542–551. IEEE.

Miao, J.; Wu, Y.; and Yang, Y. 2022. Identifying Visible Parts via Pose Estimation for Occluded Person Re-Identification. *IEEE Trans. Neural Networks Learn. Syst.*, 33(9): 4624–4634.

Ristani, E.; Solera, F.; Zou, R. S.; Cucchiara, R.; and Tomasi, C. 2016. Performance Measures and a Data Set for Multi-target, Multi-camera Tracking. In *ECCV Workshops (2)*, volume 9914 of *Lecture Notes in Computer Science*, 17–35.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 815–823. IEEE Computer Society.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *ICCV*, 618–626. IEEE Computer Society.

Shi, Y.; Ling, H.; Wu, L.; Zhang, B.; and Li, P. 2022. Attribute disentanglement and registration for occluded person re-identification. *Neurocomputing*, 470: 226–235.

Sun, Y.; Xu, Q.; Li, Y.; Zhang, C.; Li, Y.; Wang, S.; and Sun, J. 2019. Perceive Where to Focus: Learning Visibility-Aware Part-Level Features for Partial Person Re-Identification. In *CVPR*, 393–402. Computer Vision Foundation / IEEE.

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018. Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline). In *ECCV (4)*, volume 11208 of *Lecture Notes in Computer Science*, 501–518. Springer.

Tan, L.; Dai, P.; Ji, R.; and Wu, Y. 2022. Dynamic Prototype Mask for Occluded Person Re-Identification. In *ACM Multimedia*, 531–540. ACM.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NIPS*, 5998–6008.

Wang, G.; Yang, S.; Liu, H.; Wang, Z.; Yang, Y.; Wang, S.; Yu, G.; Zhou, E.; and Sun, J. 2020. High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-Identification. In *CVPR*, 6448–6457. Computer Vision Foundation / IEEE.

Wang, T.; Liu, H.; Song, P.; Guo, T.; and Shi, W. 2022a. Pose-Guided Feature Disentangling for Occluded Person Re-identification Based on Transformer. In *AAAI*, 2540–2549. AAAI Press.

Wang, Z.; Zhu, F.; Tang, S.; Zhao, R.; He, L.; and Song, J. 2022b. Feature Erasing and Diffusion Network for Occluded Person Re-Identification. In *CVPR*, 4744–4753. IEEE.

Wu, S.; Chen, Y.; Li, X.; Wu, A.; You, J.; and Zheng, W. 2016. An enhanced deep feature representation for person re-identification. In *WACV*, 1–8. IEEE Computer Society.

Yan, C.; Pang, G.; Jiao, J.; Bai, X.; Feng, X.; and Shen, C. 2021. Occluded Person Re-Identification with Single-scale Global Representations. In *ICCV*, 11855–11864. IEEE.

Yan, G.; Wang, Z.; Geng, S.; Yu, Y.; and Guo, Y. 2023. Part-Based Representation Enhancement for Occluded Person Re-Identification. *IEEE Trans. Circuits Syst. Video Technol.*, 33(8): 4217–4231.

Yang, J.; Zhang, J.; Yu, F.; Jiang, X.; Zhang, M.; Sun, X.; Chen, Y.; and Zheng, W. 2021. Learning to Know Where to See: A Visibility-Aware Approach for Occluded Person Re-identification. In *ICCV*, 11865–11874. IEEE.

Zhai, Y.; Lu, S.; Ye, Q.; Shan, X.; Chen, J.; Ji, R.; and Tian, Y. 2020. AD-Cluster: Augmented Discriminative Clustering for Domain Adaptive Person Re-Identification. In *CVPR*, 9018–9027. Computer Vision Foundation / IEEE.

Zhao, C.; Lv, X.; Dou, S.; Zhang, S.; Wu, J.; and Wang, L. 2021. Incremental Generative Occlusion Adversarial Suppression Network for Person ReID. *IEEE Trans. Image Process.*, 30: 4212–4224.

Zhao, C.; Qu, Z.; Jiang, X.; Tu, Y.; and Bai, X. 2023. Content-Adaptive Auto-Occlusion Network for Occluded Person Re-Identification. *IEEE Trans. Image Process.*, 32: 4223–4236.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable Person Re-identification: A Benchmark. In *ICCV*, 1116–1124. IEEE Computer Society.

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in Vitro. In *ICCV*, 3774–3782. IEEE Computer Society.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, 13001–13008.

Zhu, K.; Guo, H.; Liu, Z.; Tang, M.; and Wang, J. 2020. Identity-Guided Human Semantic Parsing for Person Re-identification. In *ECCV (3)*, volume 12348 of *Lecture Notes in Computer Science*, 346–363. Springer.

Zhuo, J.; Chen, Z.; Lai, J.; and Wang, G. 2018. Occluded Person Re-Identification. In *ICME*, 1–6. IEEE Computer Society.