

Feature Erasing and Diffusion Network for Occluded Person Re-Identification

Zhikang Wang^{1,2*}, Feng Zhu^{2†}, Shixiang Tang³, Rui Zhao^{2,4}, Lihuo He^{5‡}, Jiangning Song¹

¹Monash University, ²SenseTime Research, ³The University of Sydney,

⁴Qing Yuan Research Institute, Shanghai Jiao Tong University, ⁵Xidian University

zkwang00@gmail.com, zhufeng@sensetime.com, stan3906@uni.sydney.edu.au, zhaorui@sensetime.com,

lihuo.he@gmail.com, jiangning.song@monash.edu

Abstract

Occluded person re-identification (ReID) aims at matching occluded person images to holistic ones across different camera views. Target Pedestrians (TP) are often disturbed by Non-Pedestrian Occlusions (NPO) and Non-Target Pedestrians (NTP). Previous methods mainly focus on increasing the model's robustness against NPO while ignoring feature contamination from NTP. In this paper, we propose a novel Feature Erasing and Diffusion Network (FED) to simultaneously handle challenges from NPO and NTP. Specifically, aided by the NPO augmentation strategy that simulates NPO on holistic pedestrian images and generates precise occlusion masks, NPO features are explicitly eliminated by our proposed Occlusion Erasing Module (OEM). Subsequently, we diffuse the pedestrian representations with other memorized features to synthesize the NTP characteristics in the feature space through the novel Feature Diffusion Module (FDM). With the guidance of the occlusion scores from OEM, the feature diffusion process is conducted on visible body parts, thereby improving the quality of the synthesized NTP characteristics. We can greatly improve the model's perception ability towards TP and alleviate the influence of NPO and NTP by jointly optimizing OEM and FDM. Furthermore, the proposed FDM works as an auxiliary module for training and will not be engaged in the inference phase, thus with high flexibility. Experiments on occluded and holistic person ReID benchmarks demonstrate the superiority of FED over state-of-the-art methods.

1. Introduction

Person Re-Identification (ReID) aims at retrieving the same pedestrians captured by different cameras with differ-

Figure 1. Illustration of pose estimation and human parsing on pedestrian images. Both models perform well on holistic and object occluded pedestrians but fail on multi-pedestrian images. Meanwhile, human parsing models have difficulty in identifying personal belongings, e.g., backpacks, and umbrellas.

ent viewpoints, lighting conditions, and locations. With the rapid development of deep learning area and publication of large-scale image and video ReID datasets, ReID methods based on deep neural networks have achieved remarkable performance [14, 19, 27, 29]. Most of these approaches assume that a holistic body of each pedestrian is available for feature extraction. However, in real-world scenarios, e.g., railway stations, schools, hospitals, and shopping malls, pedestrians are inevitably disturbed by non-pedestrian occlusions (NPO) and non-target pedestrians (NTP). Therefore, designing a powerful network for the occluded person ReID is essential.

Methods assisted by human key points [5, 21] and human parsing information [15] dominate the state-of-the-art performance of the occluded ReID task. Generally, an auxiliary model extracts the body information first, and then the extracted information will assist the training of models. The strategy can greatly avoid mistakenly treating NPO as human parts. However, such methods have many caveats. Firstly, due to the domain gap between the training and testing data, the performance of the auxiliary models can not be consistent. In Fig. 1, we adopt of cial pose estimation

* Zhikang Wang did this work as an intern in SenseTime Research.

† Corresponding author.

‡ This research was supported partially by the National Natural Science Foundation of China (Grant Nos. 61876146).

model [22] and retrained human parsing model [37] to extract body information. It is clear that both models perform well on holistic and object occluded pedestrian images but fail on multi-pedestrian ones, which means that noise from Occluded-DukeMTMC and Occluded-REID dataset, our NTP will contaminate the global representations. Compared with object occlusion, the characteristics of NTP will result in a higher mismatching probability because of the semantic guidance. Secondly, the human parsing model can not recognize some person belonging to bags, backpacks, umbrellas, which may lead to the deficiency of valuable information. At last, the enormous computation brought by the auxiliary models makes it unacceptable for real-time video surveillance.

2. Related Works

In this section, we briefly overview the existing methods of holistic person ReID and occluded person ReID.

2.1. Holistic Person ReID

Person re-identification (ReID) aims to retrieve a person of interest in other camera views and has great progress been made in recent years. Existing ReID methods can be summarized into three categories, including hand-crafted methods [20, 34], metric learning methods [3, 41], and deep learning methods [23, 28, 32]. Due to the publishing of large-scale datasets and the development of Graphics Processing Unit (GPU), deep learning based methods have become dominant in the person re-identification area nowadays. Recent works utilizing part-based features have achieved state-of-the-art performance for the holistic person ReID. Zhan et al. [36] perform an automatic part feature alignment through the shortest path loss during the learning, without requiring extra supervision or explicit pose information. Sun et al. [23] propose a general part-level feature learning method, which can accommodate various part partitioning strategies. The attention mechanism has also been adopted to ensure the model focus on human areas, which extracts more effective features [16, 24, 33]. However, these methods fail to retrieve persons with high accuracy when occlusions happen. The shortcoming limits the utility of the methods, especially in the common crowd scenes.

2.2. Occluded Person ReID

The study of the occluded person ReID is proposed by Zhou et al. [43]. The training set and gallery set are constructed by holistic pedestrian images, and the query set is constructed by occluded pedestrian images. Recent study methods in this topic can be divided into two categories: assisted by pose estimation [10, 12] and human parsing [15, 35]. Gao et al. propose a Pose-guided Visible Part Matching (PVPM) method that jointly learns the discriminative features with pose-guided attention and self-supervised training. Li et al. [10] introduce a novel method named Pose-Guided Feature Alignment (PGFA), exploiting pose landmarks to disentangle the useful information from the occlusion noise. At the same time, extensive experiments on both occluded datasets (Occluded-DukeMTMC [21], Partial-REID [33], on human parsing. By extracting features from semantic

In summary, we propose the feature erasing and diffusion network (FED) to tackle the distractions from NPO and NTP for occluded person ReID. FED consists of three innovative components: NPO augmentation strategy, occlusion erasing module (OEM), and feature diffusion module (FDM). These components enable the network to precisely perceive the TP regardless of the NPO and NTP. At the same time, extensive experiments on both occluded datasets (Occluded-DukeMTMC [21], Partial-REID [33], on human parsing. By extracting features from semantic

Figure 2. Overview of the proposed feature erasing and diffusion network for occluded person re-identification. The two branches share the same parameters and the network consists of the feature extractor, occlusion erasing module (OEM), and feature diffusion module (FDM). The 'NPO Aug' indicates the NPO augmentation strategy. The solid lines connected to the Memory Banks indicate that the features participate in the memory update and loss calculation. The dashed lines indicate only loss calculation. The FDM is an auxiliary module for simulating NTP on feature level and will not be engaged in the inference phase.

part regions and performing comparisons with consideration of visibility, the method not only reduces background noise but also achieves body alignment.

Different from the above methods, our approach does not rely on extra models and can be trained in an end-to-end fashion. We simulate NPO and NTP on both image and feature levels and thus greatly improve the model robustness.

3. Feature Erasing and Diffusion Network

In this section, we introduce the proposed feature erasing and diffusion network (FED) in detail. The overall architecture of the network is illustrated in Fig.2. It begins with the NPO augmentation strategy that produces image pairs and occlusion masks. Following [13], we simply adopt the Vision Transformer (ViT) [4] as the feature extractor. Position embeddings and a classification [cls] token are prepended to the input image. The output feature for each image is $f \in \mathbb{R}^{(n+1) \times c}$, where $n+1$ indicates the images tokens and one [cls] token, and c is the channel dimension. Under our settings, n and c are 128 and 768, respectively. Next, we conduct the part pooling operation on image tokens and obtain N local features, which will be fed into the occlusion erasing module (OEM). Here, we set N as 4 in accordance with NPO augmentation strategy. Two memory banks will be initialized at the beginning and updated with training processing. The auxiliary feature diffusion module (FDM) takes the image features and the first memory bank as input for multi-pedestrian simulation. Details of each module will be presented in the following section.

3.1. NPO Feature Erasing

Typically, NPO feature erasing needs auxiliary information for guidance. In this section, we propose the NPO augmentation strategy and occlusion erasing module to explicitly learn NPO-robust features.

NPO Augmentation Strategy. Occlusion augmentation strategies are effective in occluded ReID. Typically, there are two categories: (1) Zhong et al. [42] randomly select a rectangle region in an image and erase its pixels with random values; (2) Chen et al. [2] paste the selected objects or backgrounds onto images. The first method helps to reduce the risk of over-fitting and makes the model robust to occlusion. However, when facing the diversified occlusions, the trained model fails to identify them due to weak generalization. The second method implicitly learns NPO-robust features by simulating the occlusion scenes. However, it fails to fully utilize the potential information, e.g., precise occlusion region, brought by the augmentation.

Inspired by the methods above, we propose the NPO augmentation strategy. The strategy consists of occlusion augmentation and mask generation, which will generate augmented images for occlusion simulation and occlusion masks for further semantic analysis, respectively.

Empirically, occlusions happen at four locations (top, bottom, left, right) with a quarter to half areas. Our augmentation strategy is similar to Chen et al. [2], but with particular modifications. For occlusion augmentation, one important step is the patch set collection. To avoid extra body parts included in the patch set, we manually crop the backgrounds and occlusion objects from the chosen images in the training set and refer to these patches as the occlusion set. We formally describe the occlusion augmentation

process as follows. Firstly, given an input image, we do common augmentations, e.g., resize, padding, and random crop, on it and get $x \in \mathbb{R}^{3 \times h \times w}$, where h and w represent the height and width, respectively. Secondly we select a patch $p \in \mathbb{R}^{3 \times p_h \times p_w}$ from the occlusion set, where p_h and p_w are the height and width. Rather than randomly paste the patch onto x , we believe that only reasonable occlusions for pedestrians can generate valuable data for training. Therefore, we calculate the aspect ratio of the patch $\alpha = p_h/p_w$. When α is larger than 3, it implies the patch is more like a vertical occlusion, otherwise horizontal occlusion. Common augmentations, e.g., random crop, and color jitter, are also applied on the patch for increasing its varieties. We resize the patch according to the occlusion type (horizontal or vertical) to $R^{(H=4 \times H=2; W)}$ and $R^{(H; W=4 \times W=2)}$, respectively. Thirdly, we randomly select one corner of x as the starting point and paste the augmented patch on it. The augmented image is named \tilde{x} .

Mask generation is a fine-to-coarse process. Firstly, we get the pixel differences by subtraction and absolute function $d = |x - \tilde{x}|$. Considering the subsequent part-based occlusion erasing module, each position of the occlusion mask should correspond to specific body parts. However, there are mis-alignments of semantics (body parts) between different images, fine-grained occlusion masks will have many false labels. Therefore, we roughly split the image into 4 stripes horizontally and aim at labeling them. As said before, there are vertical and horizontal occlusions in real-world scenarios. Vertical occlusion only damages parts of the symmetric characteristics. Usually, ReID models can easily distinguish between pedestrians and vertical occlusions and get discriminative representations without referring to further information. Therefore, the vertical occlusion is ignored while mask generation and stripes are regarded as a human part (valued 1). For the horizontal occlusion augmentation, we conduct the soft binarization operation. We take stripes covered more than three-quarters as occlusions (value 0), otherwise as human parts (value 1). In this way, we get the precise occlusion masks for the image pair.

Occlusion Erasing Module. Although the augmentation strategy is employed while training, the NPO may still contaminate representations. To further eliminate the influence of NPO, we propose the occlusion erasing module (OEM) for part feature erasing. As shown in Fig. 2, the module is constructed by 4 sub-modules corresponding to each body part. For each sub-module, it is constructed by two fully connected (FC) layers, one layer normalization [1], and one Sigmoid function. The layer normalization is placed between the FC layers, and Sigmoid function is located at the end. The first FC layer compresses the channel dimension to the quarter of the original one, aiming to wipe off the characteristic information and reserve the

Figure 3. Illustration of the feature diffusion module. The module diffuses characteristics of memory bank to the features f_i^0 for simulating NTP on feature level.

semantic ones. The next Sigmoid function will output the regressed occlusion scores s_i for each part feature. We refer to the multiplication between the occlusion scores and part features as f_i^0 . Functionally the process can be represented by

$$f_i^0 = \text{Sigmoid}(W_{rg} \text{LN}(W_{cp} f_i)) \cdot f_i; \quad (1)$$

where $W_{cp} \in \mathbb{R}^{c \times 4}$, $W_{rg} \in \mathbb{R}^{1 \times c \times 4}$, LN is the layer normalization and i indicates i -th part feature.

Here, the occlusion masks from the NPO augmentation strategy are adopted to supervise the training of OEM. We calculate the Mean Square Error (MSE) Loss between occlusion masks and occlusion scores, and the function can be expressed as

$$L_{MSE} = \frac{1}{N} \sum_{i=1}^N (s_i; \text{mask}_i); \quad (2)$$

3.2. Feature Diffusion Module

Previous works have not focused on the challenges of NPO. Apart from destroying the feature integrity of the TP, NTP also contaminates representations with realistic semantic noise. To solve this issue, we propose a learnable structure named feature diffusion module (FDM) to simulate multi-pedestrian images in the feature space. By optimizing the diffused features, we aim at indirectly enhancing the model's perception ability towards TP and robustness towards NTP. As shown in Fig. 3, apart from the image features, an extra memory bank, which is a collection of

characteristics, is taken as the input. In the following section, we will introduce OEM and FDM, respectively.

Memory Bank. The generation of M includes memory initialization and memory update. We follow the same strategy as [7]. The memory is initialized with the ID centers in the training set. We get the extracted features by performing forward computation, and average features with identical identities to get ID centers. Note that the memory initialization is only operated at the beginning of the algorithm and memory update is processed at each iteration in each mini-batch during training. The center c_k is updated by the mean of the encoded features belonging to identity k in the mini-batch as:

$$c_k = m c_k + (1 - m) \frac{1}{|B_k|} \sum_{f_i^0 \in B_k} f_i^0, \quad (3)$$

where B_k denotes the feature set belonging to identity k in the mini-batch, m is the momentum coefficient for updating, f_i^0 is the attended features after OEM. Apart from acting as the characteristic set, the memory bank is also adopted for calculating the Contrastive Loss which will be introduced in the following section. We set α as 0.2 in our experiments.

Feature Diffusion Module. Essentially, FDM is a modified cross attention module based on the standard architecture of the transformer [25]. Given the feature vector, queries Q arise from the f_i^0 , and keys K and values V arise from the memory bank M . The input feature is $f_i^0 \in \mathbb{R}^{1 \times (N \times c)}$, where N corresponds to the previous part pooling operation and is 4. Firstly, we conduct Memory Searching Operation between f_i^0 and M . It finds K nearest centers $M^K \in \mathbb{R}^{K \times (N \times c)}$ with different identities from the input image. Cosine distance is adopted for measurement. Here, we discard the center with an identical identity for avoiding polarization of the attention matrix which is calculated through cross-product. Formally,

$$Q = f_i^0 W^1; K_i = M_i^K W^2; V_i = M_i^K W^3; \quad (4)$$

where $i \in \{1; 2; \dots; K\}$, and $W_1 \in \mathbb{R}^{d^0 \times d^0}$, $W_2 \in \mathbb{R}^{d^0 \times d^0}$, $W_3 \in \mathbb{R}^{d^0 \times d^0}$ are linear projections. Then we calculate the attention matrix and corresponding part features. Formally,

$$m_i = \frac{\exp(\alpha \langle Q, K_i \rangle)}{\sum_{j=1}^K \exp(\alpha \langle Q, K_j \rangle)}; \quad \alpha = \frac{Q K_i}{d_k}; \quad (5)$$

where $\frac{1}{d_k}$ is a scaling factor. Each element of the attention matrix indirectly indicates the connections between Q and K_i , and the cross-product operation between Q and the attention matrix will generate the diffused features. The aggregation process can be defined as:

$$f_d = \text{Att}(Q; K; V) = \sum_{i=1}^K m_i V_i; \quad (6)$$

The multi-head attention operation is of great significance in this module. Since M^K has many similar patterns with the input image and these patterns are distributed randomly in K feature centers. The multi-head operation will split each center into multi parts and generate attention weights for each part individually, thus ensuring more patterns similar to TP and sufficient unique patterns of NTP can be aggregated. In this way, we can simulate the multi-pedestrian images on feature level. After the cross attention operation, we utilize the post-layer normalization feed forward network (FFN_1) [31] to conduct non-linear transformation. $\text{FFN}_1(\cdot)$ is a simple neural network with two fully connected layers and one activation function. The residual connection before the layer normalization is applied. Next, the occlusion scores generated by OEM are adopted for weighted summation between the transformed features and f_i^0 . This ensures the characteristics of NTP are only added on human parts rather than pre-recognized object occlusion parts, improving the realness and quality of the diffused features. Besides, the weighted residual operation can stabilize the training process. Then, we utilize another FFN_2 [31] for generating the final diffused representation of each image. Formally,

$$f_d^0 = \text{FFN}_2(\text{mask} \cdot \text{FFN}_1(f_d) + f_i^0); \quad (7)$$

where FFN_2 has the same structure as FFN_1 .

Since the FDM is just an auxiliary module for simulation during training, it will be removed in the inference phase. This makes our model more concise and flexible.

3.3. Loss Functions

There are three varieties of loss functions in our method, including Mean Square Error (MSE) Loss, Cross Entropy Loss, and Contrastive Loss. We refer to Cross Entropy Loss as ID Loss in this paper. As shown in Fig. 2, we calculate the ID Loss on the output features of the classification [cls] token, attended features after the OEM, and features after the FDM. Therefore, there are three additional fully connected layers on the top of the features to calculate the ID probabilities. Functionally, ID Loss can be presented as:

$$L_{ID} = - \sum_{i=1}^{ID_s} y_i \log \left(\frac{\exp(W_i f_i)}{\sum_{j=1}^{ID_s} \exp(W_j f_j)} \right); \quad (8)$$

where W is a linear projection matrix, y_i is the corresponding label and ID_s is the total number of identities. As for the Contrastive Loss, the key components are the negative and positive samples. There are two memory banks in our algorithm, the first is generated after the OEM and the second is generated after FDM. The initialization and update strategies have been introduced in Sec 3.2. Functionally, the Contrastive Loss is:

$$L_C = - \log \frac{\exp(\langle f; c_i \rangle)}{\sum_j \exp(\langle f; c_j \rangle)}; \quad (9)$$

where τ is a predefined temperature parameter and μ represents the feature center with an identical identity. Although the training strategy is a parallel architecture, the lower branch does not involve in the memory initialization and update due to the characteristic deficiency caused by the NPO augmentation. In Fig.2, we utilize the solid lines to represent jointly memory update and loss calculation and dashed lines to represent loss calculation only.

Therefore, the final loss function can be expressed:

$$L_{\text{Final}} = \frac{1}{2} \sum_{i=1}^X L_{\text{MSE}}^i + \frac{1}{2} \sum_{i=1}^X L_{\text{ID}}^i + \frac{1}{2} \sum_{i=1}^X L_{\text{C}}^i: (10)$$

4. Experiments

4.1. Datasets and Evaluation Setting

Occluded-DukeMTMC [21] 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 occluded person ReID datasets due to the diverse scenes and distractions.

Occluded-REID [43] is an occluded person ReID dataset captured by mobile cameras. It consists of 2,000 images belonging to 200 identities. Each identity has five full-body person images and five occluded person images with different viewpoints and different types of severe occlusions.

Partial-REID [39] is a specially designed ReID dataset that consists of occluded, partial, and holistic pedestrian images. It involves 600 images of 60 persons. We take the occluded query set and holistic gallery set for experiments.

Market-1501 [38] is a famous holistic person ReID 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. Few images in this dataset are occluded.

DukeMTMC-reID [40] consists of 16,522 training images of 702 persons, 2,228 queries of 702 persons, and 17,661 gallery images of 702 persons. The images are captured by 8 different cameras, making it more challenging. As it contains more holistic images than occluded ones, this dataset can be treated as a holistic ReID dataset.

Evaluation Protocol. To guarantee a fair comparison with existing person ReID methods, all methods are evaluated under the Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP). All experiments are performed in the single query setting.

4.2. Implementation Details

Unless otherwise specified, all images are resized to 256 × 128. We train our network in an end-to-end fashion through the SGD optimizer with a momentum of 0.9 and weight decay of 1e-4. We initialize the learning rate

Method	O-Duke		O-REID		P-REID	
	R@1	mAP	R@1	mAP	R@1	mAP
PCB [23]	42.6	33.7	41.3	38.9	66.3	63.8
RE [42]	40.5	30.0	-	-	54.3	54.4
FD-GAN [6]	40.8	-	-	-	-	-
DSR [9]	40.8	30.4	72.8	62.8	73.7	68.07
SFR [11]	42.3	32	-	-	56.9	-
FRR [12]	-	-	78.3	68.0	81.0	76.6
PVPM [5]	47	37.7	70.4	61.2	-	-
PGFA [21]	51.4	37.3	-	-	69.0	61.5
HOReID [26]	55.1	43.8	80.3	70.2	85.3	-
OAMN [2]	62.6	46.1	-	-	86.0	-
PAT [17]	64.5	53.6	81.6	72.1	88.0	-
ViT Baseline [13]	60.5	53.1	81.2	76.7	73.3	74.0
TransReID [13]	64.2	55.7	70.2	67.3	71.3	68.6
FED (Ours)	68.1	56.4	86.3	79.3	83.1	80.5
FED* (Ours)	67.9	56.3	87.0	79.4	84.6	82.3

Table 1. Performance comparison with state-of-the-art methods on Occluded-DukeMTMC, Occluded-REID and Partial-REID datasets. * indicates combining OS₁ and OS₂ for NPO augmentation.

Model	Market-1501		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
PT [18]	87.7	68.9	78.5	56.9
PGFA [21]	91.2	76.8	82.6	65.5
PCB [23]	92.3	77.4	81.8	66.1
OAMN [2]	92.3	79.8	86.3	72.6
BoT [19]	94.1	85.7	86.4	76.4
HOReID [26]	94.2	84.9	86.9	75.6
PAT [17]	95.4	88.0	88.8	78.2
ViT Baseline [13]	94.7	86.8	88.8	79.3
TransReID [13]	95.0	88.2	89.6	80.6
FED (Ours)	95.0	86.3	89.4	78.0

Table 2. Performance comparison with state-of-the-art methods on Market-1501 and DukeMTMC-reID datasets.

as 0.008 with cosine learning rate decay. For each input branch, the batch size is 64, which contains 16 identities and 4 samples per identity. We conduct all experiments on two RTX 1080Ti GPUs. We set the temperature in Contrastive Loss as 0.05 and the number of heads in the FDM as 8. For the occlusion set of NPO augmentation, we crop 30 patches from the training data of Occluded-DukeMTMC and MSMT17 [30] as occlusion set OS₁ and occlusion set 2 (OS₂), respectively. If not specified, we only adopt OS₁ for NPO augmentation.

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

Comparisons on Occluded Datasets. The results on Occluded-DukeMTMC (O-Duke), Occluded-REID (O-REID), and Partial-REID (P-REID) are shown in Table 1. Since O-REID and P-REID don't have corresponding train-

ing set, we simply adopt the model trained on Market-1501 for testing. PAT [17] makes great improvement on accuracy. They adopt ResNet50 [8] as the backbone and conduct diverse part discovery through the transformer encoder-decoder structure. The prototypes in the network are like specific feature detectors, which are important to improve the network performance on occluded data. TransReID [13] is the first pure transform-based architecture for person ReID. For a fair comparison, we present the results of TransReID that adopts the Vision Transformer [4] without the sliding window setting as the backbone and images resized to 256 × 128. Since He et al. [13] do not provide performance on O-REID and P-REID datasets, we retrain their official code on Market-1501 dataset and test on the two occluded datasets. The ViT Baseline performs better than TransReID on O-REID and P-REID datasets, this is because TransReID employs many dataset-specific tokens, which reduces the model's cross-domain generalization and increases the overfitting risk.

When comparing our FED (augmented OS₁) with state-of-the-art methods, we achieve the highest Rank-1 and mAP on both O-Duke and O-REID datasets. Especially on O-REID dataset, we achieve 86.3%/79.3% on Rank-1/mAP, surpassing others by at least 4.7%/2.6%. On O-Duke, we achieve 68.1%/56.4% on Rank-1/mAP, surpassing others by at least 3.6%/0.7%. On the P-REID dataset, we achieve the highest mAP accuracy, reaching 80.5% and surpassing other methods by 3.9%. We fail to achieve the highest Rank-1 accuracy on this dataset due to the low generalization of ViT backbone trained on a small dataset. Meanwhile, to further demonstrate the flexibility and scalability of the FED, we add more diversified patches (combining OS₁ and OS₂) for NPO augmentation. As we can see from the table, FED* improves Rank-1/mAP on O-REID and P-REID by at least 0.7% by simply improving the diversity of the occlusion set. In conclusion, we achieve great performance on the occluded ReID datasets.

Comparisons on Holistic Datasets We also experiment on holistic person ReID datasets, including Market-1501 and DukeMTMC-reID. While training on the DukeMTMC-reID dataset, MSE Loss is not calculated. It is because huge amounts NPO exist in the training set and we are unable to get precise occlusion masks. The results are shown in Table.2. We achieve comparable performance compared with other state-of-the-art methods. The same section 4.3.1, the TransReID is without the sliding window setting and with 256 × 128 image size. It is clear that TransReID gets better performance than our method on the holistic datasets. This is because TransReID is specially designed for holistic ReID and encodes camera information during the training process. Besides, our proposed three components, which aim at tackling the occlusion issues, are not fully functional on holistic ReID datasets. However, we also

Figure 4. Occlusion scores of OEM on horizontal occluded, vertical occluded and multi-pedestrian images. The OEM has the capacity to identify crucial NPO and fails on NTP.

Occluded-DukeMTMC							
Index	RE	NPO Aug	OEM	FDM	R@1	mAP	
0	%	%	%	%	59.1	49.1	
1	!	%	%	%	60.3	53.1	
2	%	!	%	%	65.4	53.5	
3	%	!	!	%	66.5	55.4	
4	%	!	%	!	67.1	55.9	
5	%	!	!	!	68.1	56.4	

Table 3. Performance analysis of each component in FED.

achieve 84.9% Rank-1 accuracy on DukeMTMC-reID, surpassing other CNN-based methods and close to TransReID.

4.4. Ablation Studies

Analysis of Each Component. In Table.3, we present the ablation studies of random erasing (RE), NPO augmentation strategy (NPO Aug), occlusion erasing module (OEM), and feature diffusion module (FDM). The indexes from 0 to 5 represent baseline, baseline + RE, baseline + NPO Aug, baseline + NPO Aug + OEM, baseline + NPO Aug + FDM and FED, respectively. All the models adopt ViT as the feature extractor. Model₀ and Model₁ are both optimized by ID Loss and Triplet Loss [14]. By comparing Model₀ and Model₁, we can see that RE [42] is effective in improving discrimination of representations, however the improvement is not comparable with our NPO Aug (4.9% higher on Rank 1). We can conclude that the augmented images through NPO Aug are realistic and valuable. By comparing Model₂ with Model₃, the proposed OEM can further improve the representations and improve mAP by 1.9% by removing the potential NPO information. By comparing Model₂ with Model₄, FDM helps the model with 1.7% and 2.4% improvements on Rank-1 and mAP. It means that optimizing the network with diffused features can greatly improve the model's perception ability towards TP. Finally,

Figure 5. Retrieval results of TransReID and our proposed FED on Occluded-DukeMTMC dataset. The top 2 rows show images with NPO and the bottom 2 rows show images with NTP.

Model	Occlude-DukeMTMC			DukeMTMC-reID			Market-1501		
	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP
K=2	67.4	78.4	55.8	89.4	94.7	77.6	95.0	98.6	85.7
K=4	67.7	79.9	56.2	89.2	94.3	78.0	94.8	98.5	86.3
K=6	67.3	79.8	56.2	88.9	94.2	77.3	94.8	98.4	86.0
K=8	68.1	79.3	56.4	89.0	94.3	77.1	94.8	98.4	85.9

Table 4. Analysis of the K in memory searching on both occluded and holistic datasets. CMC curve and mAP are presented for evaluation.

FED achieves the highest accuracy, demonstrating that each component can work individually and cooperatively.

Analysis of the K in Memory Searching. Here, we analyze the searching number in the memory searching operation. In Table.4, we set K as 2, 4, 6, and 8 and conduct experiments on both holistic and occluded datasets. As we can see, the performance on holistic ReID datasets appears stably on the various Ks, with a float within 0.5%. For the Market-1501, there are few NPO and NTP, failing to highlight the effectiveness of FDM. For the DukeMTMC-reID, huge amounts of training data are with NPO and NTP, and loss constraints can enable the network with high accuracy. As for the Occluded-DukeMTMC, since all the training data are holistic pedestrians, the introduction of FDM can greatly simulate the multi-pedestrian conditions in the testing set. With increasing K, FDM can better maintain the characteristics of TP and introduce realistic noise.

4.5. Qualitative Analysis

In this section, we present qualitative experimental results and demonstrate the superiority of our proposed FED.

In Fig.4, we present the occlusion scores from OEM for some pedestrian images. Images with NPO and NTP are presented. As can be seen, vertical object occlusions (Fig.4 a, b) can hardly affect the occlusion scores, since occluding less than half of symmetric pedestrians is not a critical issue for person ReID. For horizontal occlusions (Fig.4 c, d), our OEM can precisely identify NPO and label them

with smaller values. For multi-pedestrian images (Fig.4 e,f), OEM identifies each stripe as valuable. Taken together, the subsequent FDM is essential for improving the model.

In Fig.5, we present the retrieval results of TransReID and our FED. The first two examples are object occluded images. It is obvious that our network has a better recognition ability on NPO and accordingly can retrieve target pedestrians precisely. Another two examples provided are the multi-pedestrian images. Our proposed FED has a stronger perception ability on TP and achieves a much higher retrieval accuracy.

5. Conclusion

In this paper, to tackle the NPO and NTP challenges for occluded person ReID, we propose a novel Feature Erasing and Diffusion network (FED). Specifically, guided by the image-level NPO augmentation strategy, the occlusion erasing module (OEM) is trained to eliminate NPO features based on the predicted occlusion scores. Subsequently, the feature diffusion module (FDM) performs feature diffusion between NPO-feature-erased pedestrian representations and memorized features, synthesizing NTP characteristics in the feature space. Jointly optimizing OEM and FDM in our proposed FED network significantly improves the model's perception ability on TP, which is demonstrated through comprehensive experiments and comparisons with state-of-the-art algorithms on various person ReID benchmarks.

References

- [1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450* 2016. 4
- [2] Peixian Chen, Wenfeng Liu, Pingyang Dai, Jianzhuang Liu, Qixiang Ye, Mingliang Xu, Qi'an Chen, and Rongrong Ji. Occlude them all: Occlusion-aware attention network for occluded person re-id. *ICCV*, pages 11833–11842, 2021. 3, 6
- [3] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. *ICVPR* pages 403–412, 2017. 2
- [4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* 2020. 3, 7
- [5] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. Pose-guided visible part matching for occluded person reid. In *CVPR*, pages 11744–11752, 2020. 1, 6
- [6] 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 *NIPS*, pages 1229–1240, 2018. 6
- [7] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-identification. *NIPS*, 2020. 5
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, pages 770–778, 2016. 7
- [9] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. *ICVPR*, pages 7073–7082, 2018. 6
- [10] Lingxiao He and Wu Liu. Guided saliency feature learning for person re-identification in crowded scenes. *ICCV*, pages 357–373. Springer, 2020. 2
- [11] Lingxiao He, Zhenan Sun, Yuhao Zhu, and Yunbo Wang. Recognizing partial biometric patterns. *arXiv preprint arXiv:1810.07399* 2018. 6
- [12] 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*, pages 8450–8459, 2019. 2, 6
- [13] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. *ICCV*, 2021. 3, 6, 7
- [14] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* 2017. 1, 7
- [15] Houjing Huang, Xiaotang Chen, and Kaiqi Huang. Human parsing based alignment with multi-task learning for occluded person re-identification. *ICME*, pages 1–6. IEEE, 2020. 1, 2
- [16] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. *CVPR*, pages 2285–2294, 2018. 2
- [17] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. *CVPR*, pages 2898–2907, 2021. 6, 7
- [18] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *CVPR*, pages 4099–4108, 2018. 6
- [19] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPRW*, pages 0–0, 2019. 1, 6
- [20] Bingpeng Ma, Yu Su, and Frederic Jurie. Covariance descriptor based on bio-inspired features for person re-identification and face verification. *Image and Vision Computing* 32(6-7):379–390, 2014. 2
- [21] 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. 1, 2, 6
- [22] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. 2
- [23] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with re-net part pooling (and a strong convolutional baseline). *ICCV*, pages 480–496, 2018. 2, 6
- [24] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. Aanet: Attribute attention network for person re-identifications. In *CVPR*, pages 7134–7143, 2019. 2
- [25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, pages 5998–6008, 2017. 5
- [26] 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. *CVPR*, pages 6449–6458, 2020. 6
- [27] Zhikang Wang, Lihuo He, Xinbo Gao, and Yuanfei Huang. Multi-scale spatial-temporal network for person re-identification. In *ICASSP*, pages 2052–2056. IEEE, 2019. 1
- [28] Zhikang Wang, Lihuo He, Xinbo Gao, and Jane Shen. Robust person re-identification through contextual mutual boosting. *arXiv preprint arXiv:2009.07491* 2020. 2
- [29] Zhikang Wang, Lihuo He, Xiaoguang Tu, Jian Zhao, Xinbo Gao, Shengmei Shen, and Jiashi Feng. Robust video-based person re-identification by hierarchical mining. *TCSVT*, 2021. 1
- [30] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018. 6
- [31] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *ICML*, pages 10524–10533. PMLR, 2020. 5

- [32] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In CVPR, pages 2119–2128, 2018. 2
- [33] Fan Yang, Ke Yan, Shijian Lu, Huizhu Jia, Xiaodong Xie, and Wen Gao. Attention driven person re-identification. Pattern Recognition 86:143–155, 2019. 2
- [34] 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. Springer, 2014. 2
- [35] Shijie Yu, Dapeng Chen, Rui Zhao, Haobin Chen, and Yu Qiao. Neighbourhood-guided feature reconstruction for occluded person re-identification. arXiv preprint arXiv:2105.07345 2021. 2
- [36] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Align-dreid: Surpassing human-level performance in person re-identification. arXiv preprint arXiv:1711.08184 2017. 2
- [37] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In CVPR, pages 2881–2890, 2017. 2
- [38] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In ICCV, pages 1116–1124, 2015. 2, 6
- [39] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In ICCV, pages 4678–4686, 2015. 6
- [40] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In ICCV, pages 3754–3762, 2017. 2, 6
- [41] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In CVPR, pages 1318–1327, 2017. 2
- [42] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In AAAI, volume 34, pages 13001–13008, 2020. 3, 6, 7
- [43] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. Occluded person re-identification. In ICME, pages 1–6. IEEE, 2018. 2, 6