Occluded Person Re-Identification with Single-scale Global Representations

Cheng Yan[1]*, Guansong Pang[2*†], Jile Jiao[3] Xiao Bai[1†], Xuetao Feng[3], Chunhua Shen[4]
[1]Beihang University [2]The University of Adelaide [3]Alibaba Group [4]Monash University

Abstract

Occluded person re-identification (ReID) aims at re-identifying occluded pedestrians from occluded or holistic images taken across multiple cameras. Current state-of-the-art (SOTA) occluded ReID models rely on some auxiliary modules, including pose estimation, feature pyramid and graph matching modules, to learn multi-scale and/or part-level features to tackle the occlusion challenges. This unfortunately leads to complex ReID models that (i) fail to generalize to challenging occlusions of diverse appearance, shape or size, and (ii) become ineffective in handling non-occluded pedestrians. However, real-world ReID applications typically have highly diverse occlusions and involve a hybrid of occluded and non-occluded pedestrians. To address these two issues, we introduce a novel ReID model that learns discriminative single-scale global-level pedestrian features by enforcing a novel exponentially sensitive yet bounded distance loss on occlusion-based augmented data. We show for the first time that learning single-scale global features without using these auxiliary modules is able to outperform the SOTA multi-scale and/or part-level feature-based models. Further, our simple model can achieve new SOTA performance in both occluded and non-occluded ReID, as shown by extensive results on three occluded and two general ReID benchmarks. Additionally, we create a large-scale occluded person ReID dataset with various occlusions in different scenes, which is significantly larger and contains more diverse occlusions and pedestrian dressings than existing occluded ReID datasets, providing a more faithful occluded ReID benchmark. The dataset is available at: https://git.io/OPReID

1. Introduction

Person re-identification (ReID) aims to search for the same person from a gallery of pedestrian images taken from different cameras, which is a critical task in computer vision due to its broad applications in multi-camera tracking, video surveillance, and forensic search. Most existing ReID approaches [7, 16, 32, 36, 43] generally assume that the whole body of the person is visible. However, in real applications, many pedestrian images can be occluded by various obstacles such as cars, trees, and crowd. It is challenging for the general ReID approaches to learn effective representations of these occluded images [6, 15, 26, 34, 37, 39], leading to ineffective performance for occluded ReID [22, 47].

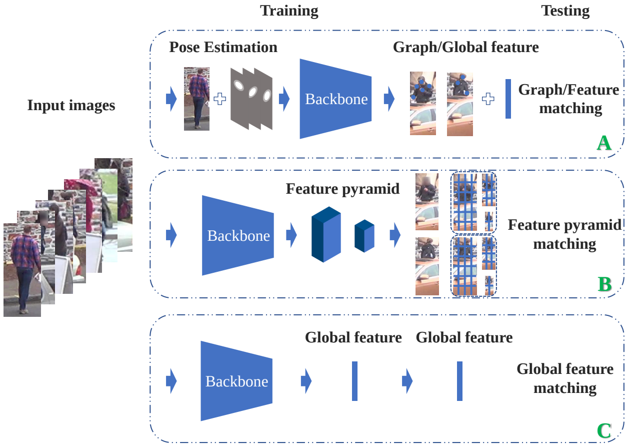


Figure 1 – Overview of three occluded ReID approaches. Existing approaches (A&B) rely on auxiliary modules, such as pose estimation and feature pyramid, to learn multi-scale or part-level features for occluded ReID, while our proposed approach (C) can address the problem with significantly simplified models that learn single-scale global features with single backbone.

A number of occluded ReID methods [8, 12, 14, 22, 33] have been proposed to tackle this problem. The current state-of-the-arts can be roughly divided into two categories, key-points based methods and feature pyramid matching methods. As shown in Figure 1 (A), the key-points based methods [8, 22, 33] often utilize pose estimation models to obtain some extra semantic information such as key-points heat-maps/graphs for identity matching in the training or inference stage. These **part-level features** or graph matching strategies help overcome the occlusion problem. However, the ReID performance is heavily relied on the performance of pose estimation models. Further, these pose es-

[*] CY and GP equally contributed to this work. CY and XB are with State Key Laboratory of Software Development Environment, Jiangxi Research Institute, Beihang University. †XB and GP are corresponding authors, e-mail: baixiao@buaa.edu.cn, pangguansong@gmail.com

timation models may also suffer from occlusions [4, 28]. As shown in Figure 1 (B), the other type of occluded ReID methods [12,14] is built upon the feature pyramid matching framework using a single backbone network during training. At the inference stage, they extract **multi-scale features** from both the query and gallery images for person matching. This feature pyramid matching strategy works well for the easy occlusions, in which only the query images contain occluded body parts and the gallery images are holistic, but it fails to deal with the cases that both the query and gallery contain occluded images, especially when the occlusions are of different appearance, shape or size. Further, without the support of key-points models, the feature pyramid matching strategy can be strongly biased by the occlusion obstacles. For example, when the query image contains similar car obstacles as in the person images in the gallery, as illustrated in Figure 1(B), these models would wrongly yield a high matching score due to the similarity of the occlusion objects rather than the person appearance. Additionally, these auxiliary modules in both types of methods can lead to over-complicated models, rendering them less effective in handling non-occluded pedestrians. However, real-world ReID applications typically have highly diverse occlusions and involve a hybrid of occluded and non-occluded pedestrians.

In this work we propose a novel ReID model that learns discriminative **single-scale global-level pedestrian representations** for such real-world ReID applications. Although it is a largely simplified model that does not require auxiliary models, it can well generalize to diverse occlusions and perform effectively in handling both occluded and holistic pedestrians. Particularly, our method leverages an occlusion-based data augmentation and an exponentially sensitive yet bounded distance loss to learn fine-grained discriminative features from non-occluded body parts. In doing so, our model is optimized in an end-to-end fashion using a single backbone network (e.g., Resnet-50), as shown in Figure 1 (C). The network is enhanced by disentangled non-local (DNL) operations and our proposed reconstructive pooling layer to better learn the non-occluded features. During inference, it uses only the single-scale global feature representations from the final feature layer, rather than the multi-scale or multiple part-level features as in current models, for person matching.

Further, there are limited publicly available dataset benchmarks for the occluded ReID task. Existing relevant datasets, including P-iLIDS [12], P-ReID [42], O-ReID [47] and O-Duke [22], are small and have too monotonous occlusions to represent the problem complexities in real-world applications. Even worse, these monotonous occlusions may mislead the design and evaluation of ReID models. For example, in the largest dataset O-Duke, most persons in the query image set are occluded by the same car at the same body parts such as legs and feet (see Figure 4), which may lead to a "Clever Hans Phenomenon" [17], e.g., ReID models can achieve 'correct' ReID based on features extracted from the occlusion objects rather than the person appearance. To address these issues, we create a large-scale occluded ReID dataset with highly diverse occlusions (e.g. carts, signs, storage racks, etc.) at different body parts from diverse scenes inside and outside the supermarkets and shopping malls (see Section 4 and Supplementary Materials). It includes 7,918 identities with all seasons' dressing and 72,442 images collected using 6,000 cameras. The resulting data is significantly more faithful and larger than the existing largest occluded ReID dataset – O-Duke that contains only 1,812 identities with exclusively winter dressing and 35,489 images collected using 8 cameras. The dataset is carefully prepared to avoid the privacy problem.

Table 1 – Modules and features used in occluded ReID methods. Modules include Single-Backbone (S-B), Pose-Estimation (P-E), Feature-Pyramid-Matching (FP-M) and Graph-Matching (G-M). Features include Multi-Scale-Feature (MS-F), Part-level-Feature (Part-F) and Global-level-Feature (Global-F).

	Module				**Feature**		
Method	S-B	P-E	FP-M	G-M	MS-F	Part-F	Global-F
DSR [12]	✓		✓		✓	✓	✓
FPR [14]	✓		✓		✓	✓	✓
PGFA [22]	✓	✓				✓	✓
PVPM [8]	✓	✓		✓		✓	✓
HOReID [33]	✓	✓		✓		✓	✓
Ours	✓						✓

In summary, our main contributions are as follows.

- We propose to tackle occluded person ReID by learning single-scale global feature representations using a single network backbone, contrasting to current state-of-the-art models that learn multi-scale and/or part-level features, with their performance heavily dependent on one or more auxiliary modules, as summarized in Table 1. To this end, we introduce a novel ReID model that minimizes an exponentially sensitive yet bounded distance loss on occlusion-based augmented data to learn such global features. Through this model, we show for the first time that learning single-scale global features outperforms the multi-scale or part-level features for occluded ReID (e.g., up to 17%-19% relative improvement in both top-1 accuracy and average precision), providing important implication for exploring simple and effective ReID models.
- We further show that our model can achieve new SOTA performance in both occluded and non-occluded ReID, as shown by extensive results on three occluded and two general ReID datasets. This is an important capability for real-world ReID applications that typically involve a hybrid of occluded and holistic pedestrians.
- We introduce a large-scale occluded person ReID

dataset with both indoor and outdoor occlusions in different scenes, which is significantly larger and contains substantially more diverse occlusions and pedestrian dressings than existing occluded ReID datasets, providing a more faithful occluded ReID benchmark.

2. Related Work

Most existing person ReID methods [7,15,16,26,32,34,36,37,39,43] focus on learning global or local features with an implicit assumption that the pedestrians are well presented in the images. These methods achieve good performance for persons with holistic appearances, but they cannot work well for occluded persons since the extracted features are misled and/or biased by the occlusions. To address this issue, there have been a number of methods introduced for occluded ReID. These methods can be roughly divided into two groups, including key-points based methods and feature pyramid/multi-feature matching-based methods.

Key-points based methods [8, 22, 33] are motivated by the success of the mask pose-guided or mask-guided methods [20, 23, 40] for the general ReID task, yet they are specifically designed to deal with the occluded ReID problem. Though these methods show effective occluded ReID performance, they heavily rely on the performance of auxiliary pose estimation models. Some recent occluded ReID methods [8, 33] show that adding extra complex graph matching modules in training or/and inference stage can help further improve the performance.

Feature pyramid matching-based methods [4,11,28,30] do not rely on auxiliary models to gain extra semantic information. The key ingredients here are the multi-feature learning and the matching between the occluded and holistic images. In [12, 14], multi-scale features are first extracted to form feature pyramids. The feature pyramids from holistic images are then used to reconstruct that of occluded images to accomplish max-matching. These methods usually work well for occluded-holistic image matching, but they fail to deal with the occluded-occluded image matching, especially in the scenarios that the occlusions are different in the query and gallery sets.

3. The Proposed Approach

3.1. Overview

Given a set of training images $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ and the corresponding one-hot encoding of the identity/class set $\mathcal{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_N\}$, our approach is to learn a single-scale global feature mapping function $\phi : \mathcal{X} \times \mathcal{Y} \mapsto \mathcal{Z}$ using one single network backbone. ϕ projects the data $\mathcal{X}$ onto a new feature space $\mathcal{Z}$, such that the distance of the images of each person is small while the distance w.r.t different persons is large. Given a query image $\mathbf{q}$, the system first computes the distance between $\phi(\mathbf{q})$ and each image $\phi(\mathbf{g}_i)$ from a gallery image set $\mathcal{G} = \{\mathbf{g}_1, \cdots, \mathbf{g}_M\}$, and then returns the images that have the smallest distance to the query image. It is generally assumed that the identities in the gallery set $\mathcal{G}$ and the training set $\mathcal{X}$ have no overlapping.

To this end, we introduce an occlusion-based data augmentation method and an exponentially sensitive yet bounded distance loss function to learn discriminative global representations of non-occluded body parts. Particularly, as shown in Figure 2, the input images are first augmented by our data augmentation method named Compound Batch Erasing (CBE), aiming to augment data as well as create synthetic occlusions in the input images, then fed into the neural network. The Disentangled Non-Local operation (DNL) is leveraged to define a novel reconstructive pooling layer in the backbone to guarantee that the network focuses on the persons rather than the occlusions. After that, the feature representations are optimized by minimizing our proposed Bounded Exponential Distance (BED) loss to attend to the discriminative non-occluded parts of the persons. During inference, only the single global representation after the global pooling is used for identity matching per query.

Additionally, we find there is a long-tailed problem in real-world ReID applications. We showed empirically that the bad momentum-effect elimination technique [31] could be used to alleviate the problem,

3.2. Non-occluded Feature Learning

Our model learns fine-grained discriminative features from non-occluded body parts by enforcing the proposed BED loss on the occlusion-based augmented data. A disentangled non-local (DNL) operation is leveraged to devise a novel reconstructive pooling layer in our backbone to further enhance the feature learning.

Occlusion-based Data Augmentation. We propose the compound batch erasing augmentation method to simulate occlusions, which includes the widely-used random erasing (RE) [45] and our proposed batch-constant erasing (BcE). These two erasing operations are separately applied to the same batch of raw pedestrian images, resulting in two augmented image batches with different image patches erased. Their combination helps produce images with occlusions at diverse body parts of different size. These two image batches are combined as one large batch to feed to the model. To achieve this, we first sample a batch of images $\mathbf{X} \in \mathbb{R}^{B\times3\times H\times W}$, we then duplicate the sub-batch and concatenate the two sub-batches to form the full batch $\mathbf{X}_{\mathit{full}} = [\mathbf{X}_{\mathit{re}}; \mathbf{X}_{\mathit{bce}}] \in \mathbb{R}^{2B\times3\times H\times W}$, of which RE and BcE are then respectively applied to $\mathbf{X}_{\mathit{re}}$ and $\mathbf{X}_{\mathit{bce}}$. This provides large-scale data for learning features for the identities with diverse blank occlusions of varying sizes.

For BcE, we erase a striping part of the image and the erased part is fixed and applied to all the images in the same sub-batch $\mathbf{X}_{\mathit{bce}}$. The erased part per sub-batch is randomly

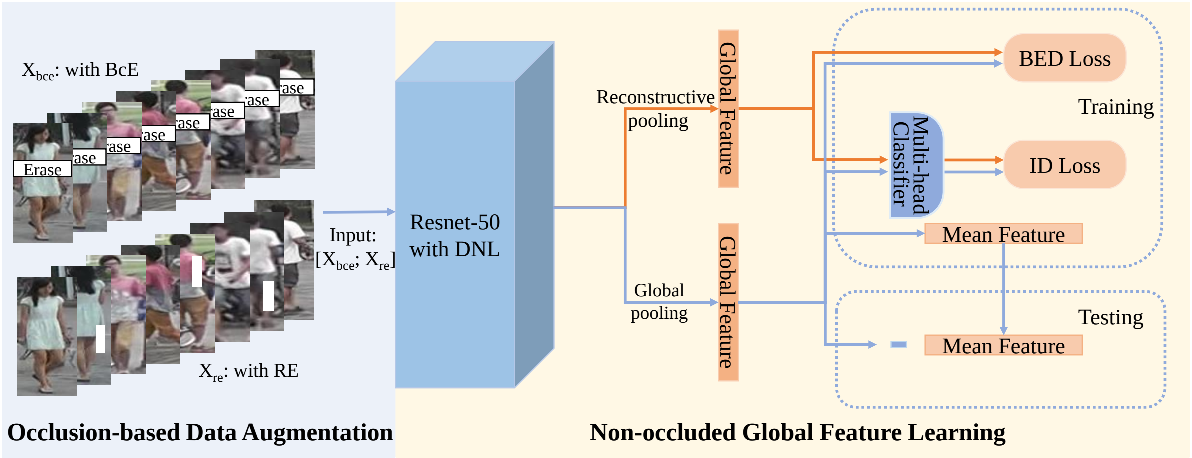


Figure 2 – An overview of the proposed approach. The compound batch erasing operation is first applied to the input images of two sub-batches - $\mathbf{X}_{re}$ and $\mathbf{X}_{bce}$ - that are made up with the same images but having RE and BcE erasing operation respectively. These two sub-batch samples are fed into a Resnet backbone with DNL blocks. At the end of the network, two feature representations are obtained by two pooling layers, the proposed reconstructive pooling and the global pooling. For training, the proposed BED loss together with the identity loss is applied to both features to learn discriminative global-level features. We also maintain a continuously updated mean feature vector to capture possible momentum bias towards large classes in the training data. During inference, the single-scale global pooling features (subtracting the mean features to alleviate the long-tailed problem) is used for person ReID.

chosen as follows. We first spatially divide the image into s horizontal parts, where $s \in \{6, 7, 8\}$ is a random value, then randomly choose one part and erase the same part of all the images per $\mathbf{X}_{bce}$. The RE part is the same as [45].

DNL-enabled Backbone and Reconstructive Pooling. Previous studies [8,12,22] show that single backbones without local feature matching or graph matching modules are ineffective in learning non-occluded features. Motivated by the tremendous success of self-attention or non-local feature learning techniques in many CV applications [2,35], we incorporate Disentangled Non-Local (DNL) operations [38] to better extract non-occluded features. This type of operations computes the neuron activation at a position as a weighted sum of the features at all positions. As a result, the features from the same *instance* (e.g., some body parts of the pedestrians) are mutually reinforced and easily attended by the model. This is an important capacity of a backbone for occluded ReID. Particularly, DNL consists of two terms, with one term accounting for the relationship between two pixels and the other one representing the saliency of every pixel. Thus, it can work effectively to obtain high response from and attend to both non-occluded body parts and salient boundaries. We follow [35] to add two/three DNL layers to the second/third stage of Resnet-50. The key formulation of pair-wise activation of DNL is as follows:

$$w(\mathbf{f}_i, \mathbf{f}_j) = s\left((\mathbf{W}^q\mathbf{f}_i - \mu_q)^T(\mathbf{W}^k\mathbf{f}_i - \mu_k)\right) + s(\mathbf{W}^m\mathbf{f}_j), \tag{1}$$

where $\mathbf{f}_i$ is a c-dimensional feature at position i of the feature map, $\mathbf{W}^q, \mathbf{W}^k \in \mathbb{R}^{c \times c}$ and $\mathbf{W}^m \in \mathbb{R}^{1 \times c}$ are linear transformation, $\mathbf{W}^q\mathbf{f}_i$ and $\mathbf{W}^k\mathbf{f}_i$ are the query and key of the non-local features, μ_q and μ_k is the mean of all queries and keys, and $s(.)$ is a softmax function. As discussed in [38], the first term is a whitened pairwise term for instance response learning and the second term focuses on salient regions. The output of position i is $\sum_j w(\mathbf{f}_i, \mathbf{f}_j) \cdot \mathbf{W}^v\mathbf{f}_i$, in which $\mathbf{W}^v \in \mathbb{R}^{c \times c}$ is another linear transformation to calculate the value term $\mathbf{W}^v\mathbf{f}_i$ in [38]. By broadcasting this to the whole feature map, we obtain the activation map in the form of a $c \times wh$ tensor. This tensor is then added to the feature map to enhance the backbone.

Further, the second term in Eq. (1) is similar to mask-based matching methods in [1, 12]. $\mathbf{W}^m\mathbf{f}_j$ is equivalent to a mask for the feature map and is used to re-weight the features. Motivated by this observation, we define a novel operation, termed **reconstructive pooling**, and incorporate it into our backbone by :

$$f_r = (\mathbf{W}^v\mathbf{F})s(\mathbf{W}^m\mathbf{F})^T, \tag{2}$$

where $\mathbf{F} \in \mathbb{R}^{c \times wh}$ is the final feature map and $\mathbf{W}^m \in \mathbb{R}^{1 \times c}$ and $\mathbf{W}^v \in \mathbb{R}^{c \times c}$ are linear transformation. The term $(\mathbf{W}^v\mathbf{F})$ is similar to the reconstruction term in the feature pyramid matching framework [12, 14], in which it is used to re-weight holistic image features to reconstruct occluded image features. $s(\mathbf{W}^m\mathbf{F})$ works as a feature selection layer on channels. For a given feature map, Eq. (2) essentially works as both feature reconstruction and pooling operations consecutively. We therefore apply Eq. (2) on top of the last feature map of the backbone to aggregate important non-occluded features. Unlike the widely-used global pooling, this pooling enables the backbone network to allocate attention to the salient boundaries as well as to reconstruct occluded features. Thus, our network itself can capture the non-occluded features without involving any part feature-based matching modules. Note that the reconstructive pooling is used only in the training stage. For inference, only the features after the global pooling is used.

Bounded Exponential Distance Loss. The occlusion-based data augmentation generates many hard examples that have synthetic occlusions. To learn discriminative features for these hard examples, we propose the novel Bounded Exponential Distance (BED) loss.

BED is a dual dynamically scaled exponential loss function, in which exponential penalization is enforced on image pairs of small distance to learn diverse local discriminative parts, enabling models to capture fine-grained difference between highly similar pedestrian images. A scaling factor is incorporated into the loss function to control the penalization sensitivity, helping automatically reduce the model's attention on easily discriminative image pairs while at the same time attending to hardly discriminative image pairs. Specifically, let $\mathbf{z}_i = \phi(\mathbf{x}_i)$ be the feature representation after the pooling layer, the BED loss is defined as follows.

$$L_{bed}(\mathbf{x}_i, \mathbf{x}_j) = y_{ij}\left(1-e^{-\alpha d(\mathbf{z}_i,\mathbf{z}_j)}\right)+(1-y_{ij})e^{-\alpha d(\mathbf{z}_i,\mathbf{z}_j)}, \tag{3}$$

where α is the sensitivity scaling factor, $d(\cdot,\cdot)$ is the ℓ_2 distance, and $y_{ij} = 1$ if $\mathbf{z}_i$ and $\mathbf{z}_j$ are from the same person and $y_{ij} = 0$ otherwise.

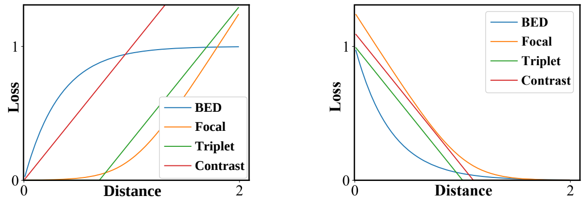
This loss has the following two key desired properties:

- It gives aggressively exponential punishment on the pairs that have small inter-/intra- person distances, enforcing attention to diverse discriminative features.
- It imposes bounded loss on image pairs of large differences, which helps automatically down-weight the easy image pairs that have large inter-/intra- person distances. The bounds are shown as follows.

$$\begin{aligned}\lim_{d_{ij}\to\infty} L_{bed}(\mathbf{x}_i, \mathbf{x}_j) = 1 \text{ if } y_{ij} = 1,\\ \lim_{d_{ij}\to\infty} L_{bed}(\mathbf{x}_i, \mathbf{x}_j) = 0 \text{ if } y_{ij} = 0,\end{aligned} \tag{4}$$

where d_{ij} presents the distance between $\mathbf{z}_i$ and $\mathbf{z}_j$.

One main benefit brought by the exponentially sensitive penalization is the capability in learning the fine-grained difference of the non-occluded body parts. Particularly, as shown in Figure 3(a), for image pairs that have small intra-person distance, our exponential penalty is different from the small or no penalization in the contrastive loss and the triplet loss. Moreover, the punishment of the triplet loss and contrastive loss is boundless, attracting the model's overwhelming attention to the image pairs with large intra-person distance. Consequently, the models based on these loss functions fail to learn the fine-grained difference of the non-occluded parts. By contrast, the BED loss has an upper bound penalization on those image pairs while enforcing exponential penalization on image pairs with small intra-person distances, enabling effective learning of the discriminative non-occluded features. In Figure 3(b), in a similar sense, for image pairs that come from different persons but have small distances, the BED loss applies penalization inversely exponential to the distance and applies nearly zero loss to the pairs that have large inter-person distance, which helps better discriminate image pairs that have small inter-person distances than the linear penalization in the contrastive loss or triplet loss.



(a) Same person (b) Different persons

Figure 3 – (a) Loss w.r.t. intra-person distance and (b) Loss w.r.t. inter-person distance. The margin in the triplet loss and contrastive loss is set to 0.3 according to [21]. One desired property of the proposed BED loss function is that its penalization is exponentially sensitive yet provably bounded. As shown in (a) and (b), our loss function exponentially punishes the similar/dissimilar pairs that have small distance and imposes bounded loss to large distances. This empowers the model to capture fine-grained discriminative features for the non-occluded body parts. By contrast, the punishment of contrastive loss, triplet loss and focal loss is substantially less sensitive and boundless for the same person, leading to an optimization that biases towards images of large intra-person distances. Consequently, the models fail to attend to scatterly distributed small discriminative parts of the images.

The exponential penalization of the BED loss is inspired by the success of the focal loss [19] in dense object detection, but the BED loss is fundamentally different from the focal loss in that (i) they have very different penalization properties (e.g., provably bounded vs. boundless), as shown in Figure 3, and (ii) the BED loss is devised to enable the learning of fined-grained features in a ranking task whereas the focal loss is introduced to address the class imbalance problem in a classification task.

The Overall Loss Function. The final loss function of our model is as follows

$$L = L_{bed}^{rpf} + L_{bed}^{gf} + L_c, \tag{5}$$

where L_{bed}^{rpf} and L_{bed}^{gf} refer to the use of the BED loss on feature representations from the reconstructive pooling and global pooling layers respectively, and L_c =

$\sum_i^H \mathbb{E}(\bar{\mathbf{Y}}_i, \mathbf{Y}_i)$ is a cross-entropy-based multi-head classification loss, which is shown to be more effective than single-head classification in [31]. Following [31], two heads, i.e., $H = 2$, are used in our experiments.

Additionally, long-tailed class distribution is commonly seen in real-world ReID applications, because there are only limited images for most of the identities while some identities have a large number of images available for training. This problem becomes more severe for occluded ReID as occlusions in the data-poor identities reduce the supervisory information for the tail classes. As a result of this class imbalance, the models can be biased towards the head classes. The bad Momentum-effect Elimination (ME) strategy [31] is used in our implementation to alleviate the problem. Particularly, the momentum in the SGD optimization can be dominated by the samples from the data-rich head classes, leading to all features inclining to the feature-directions of the head classes. We therefore maintain a continuously updated mean feature vector to capture this momentum shift towards the head classes. The mean feature vector is obtained from the features after the global pooling. It is incrementally updated by:

$$\mu_t = \beta\mu_{t-1} + (1-\beta)\cdot \widetilde{f}_g, \tag{6}$$

where μ_{t-1} is the updated mean feature vector obtained at iteration $t-1$, $\widetilde{f}_g = \frac{1}{|\mathcal{B}|}\sum_{\mathbf{z}\in\mathcal{B}} \mathbf{z}$ denotes the mean feature vector of all samples in the batch $\mathcal{B}$ at the current iteration, and β is a hyper-parameter to balance the importance of μ_{t-1} and $\widetilde{f}_g$. Since the features are all normalized after the global pooling, subtraction by this mean vector can effectively take away the bias towards the head classes. Therefore, during inference, feature representations of the images the query and gallery sets are subtracted by this mean feature vector before performing image matching.

4. The Proposed Dataset: OPReID

The proposed Occluded Person ReID (OPReID) benchmark is created based on a dataset collected from 6,497 different cameras of 89 camera systems inside and outside the supermarkets and shopping malls. The original dataset contains 1.38 million images of over 30K different identities. We use pose estimation models to classify whether an image contains an occluded or non-occluded identity. These occluded or non-occluded image classification results are then manually examined and corrected. We further randomly select occluded and non-occluded images from this large-scale dataset to create our OPReID dataset. Specifically, we first use the pose estimation method in [3] to detect key-points of all images. The images that miss at least three sequential key-points are then considered as occluded images. We select 4,200 images from 3,515 identities to form the query set. Since in real-world ReID application we often have very limited images for the identity, we collect only two images coming from the same identity of each query, including one occluded image and one holistic image, to compose a subset of the gallery set. We then randomly select 1-3 occluded images and 1-3 holistic images per identity from other identities to form the rest of the gallery, resulting in 41,014 images from 6,174 identities in the gallery set. For the training set, it contains 27,228 images from 1,744 identities, with around 15% occluded images and 85% randomly selected holistic images; there are 5-20 images for each identity. All the identities from the training set are different from that of the gallery set.

A comparison of the key statistics and characteristics between OPReID and existing occluded ReID datasets is shown in Table 2. Among these existing datasets, P-iLIDS, P-ReID and O-ReID do not have training data, and they have only very small query and gallery sets. O-Duke is a large dataset, but the occlusions in this dataset are very monotonous, e.g., most persons are occluded by the same car at the same body parts such as legs and feet (see the fourth row in Figure 4). Compared to these datasets, OPReID provides a significantly more realistic occluded ReID testbed in that (i) it contains significantly more occlusions of diverse appearance, shape and size, (ii) it contains significantly larger number of identities and images collected using 6,000 cameras in all four seasons, and (iii) it is collected from both indoor and outdoor scenes. Further, in datasets like O-Duke, too monotonous occlusions may lead to a “Clever Hans Phenomenon”, e.g., ReID models can perform ‘correct’ image matching due to solely the similarity of the occlusion objects rather than the person appearance. By contrast, our dataset contains a wide range of occlusions at different body parts, such as carts, elevators, storage racks, etc. from diverse scenes inside and outside the supermarkets and shopping malls (see the bottom row in Figure 4). All faces in OPReID are masked for privacy protection. The data is available at `https://git.io/OPReID`.

5. Experiments

5.1. Implementation Details and Datasets

The implementation of our method is built upon the FastReID-strong-baseline method in [13], in which Gem pooling [24] and non-local [35] blocks are adopted to enhance the state-of-the-art model in [21]. For fair comparison, Resnet-50 is used as backbone in which the stride of last CNN layer is set to 1. For training, the batch size, α, and β are respectively set to 64, 0.3 and 0.9 by default.

We evaluate the performance on three occluded person ReID datasets, include two popular large benchmarks, Occluded-Duke (O-Duke) [22] and Occlude-ReID (O-ReID) [47], and our dataset OPReID. Our method is also evaluated on two widely-used general person ReID datasets,

Table 2 – Characteristics of OPReID vs. existing occluded ReID datasets. Imgs and Cams are short for images and cameras, respectively.

Dataset	Train		Query		Gallery		Overall		Key Characteristic			
	#IDs	#Imgs	#IDs	#Imgs	#IDs	#Imgs	#IDs	#Imgs	Cams	Occlusion	Occluded Part	Dressing
P-iLIDS [12]	-	-	119	119	119	119	119	119	2	truncated images to simulate occlusions		winter
P-ReID [42]	-	-	60	300	60	300	60	300	5	truncated images to simulate occlusions		summer
O-ReID [47]	-	-	200	1,000	200	1,000	200	1,000	5	wall, garbage-can, umbrella	all body parts	all seasons
O-Duke [22]	702	15,618	519	2,210	1,110	17,661	1,812	35,489	8	car, billboard	leg, feet	winter
OPReID	1,744	27,228	3,515	4,200	6,174	41,014	7,918	74,416	6,000	cart, elevator, storage rack, etc.	all body parts	all seasons

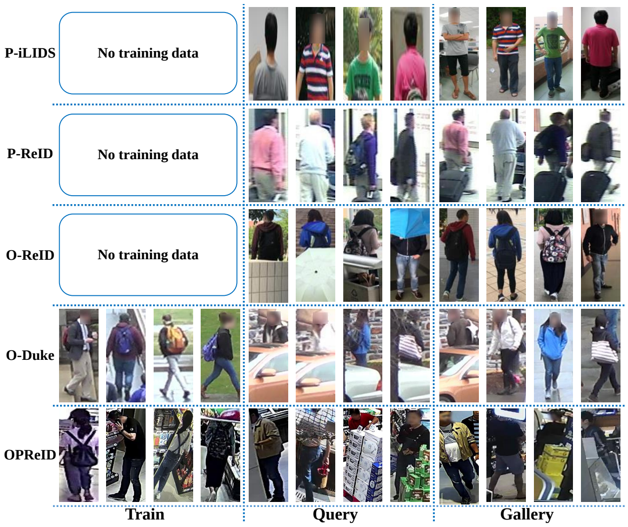


Figure 4 – Visual comparison of OPReID and existing occluded ReID datasets. We mask all faces for privacy protection.

Market1501 [41] and DukeMTMC-ReID [44]. Following [8, 21, 22], we use Rank-1 accuracy (R-1) [10] and mean Average Precision (mAP) [41] as the performance metrics.

Readers are referred to Supplementary Materials for results using other backbones, R-5, R-10 and runtime results.

5.2. Effectiveness on Occluded ReID Datasets

Our model is compared with seven state-of-the-art occluded ReID methods [8,9,12,14,16,22,33] on the three occluded ReID datasets. As O-ReID does not have a training set, we follow the literature [12,33] to train the model using the Market1501 data. We also compare our model with five state-of-the-art general ReID methods [13, 21, 26, 27, 46]. The comparison results are shown in Table 3. Note that on the proposed OPReID dataset, we report the results of three recent occluded ReID methods [12, 22, 33] and one general ReID method [13] that we obtained using their publicly available codes.

Our model substantially outperforms all the competing methods in both mAP and R-1 on two challenging datasets O-Duke and OPReID that have more complex occlusions than O-ReID. Particularly, our model outperforms the state-of-the-arts by over 13% in both R-1 and mAP on O-Duke and 17%-19% in both R-1 and mAP on OPReID. Compared with the current state-of-the-art general ReID model, Baseline [13], our model still achieves 4.7%-6.2% and 4.4%-5.3% improvement in R-1 and mAP, respectively.

On the small dataset O-ReID, incorporating heuristics/domain knowledge into ReID models or image matching is often more effective than data-driven representation learning. As a result, HOReID [33] that combines auxiliary pose information using graph convolutional networks in its image matching obtains better R-1 than our model. However, HOReID is less effective than our model in mAP on O-ReID, because the extracted pose information in different images of the same identity can be highly dissimilar, leading to a lower recall. Further, this type of methods fails to work on large-scale datasets O-Duke and OPReID, because they are difficult to specify well generalized heuristics/domain knowledge in such cases.

5.3. Effectiveness on General ReID Datasets

Motivated by the fact that real-world ReID applications typically involves a hybrid of occluded and holistic pedestrian images, our model is also evaluated on general ReID datasets to examine its applicability in handling non-occluded ReID cases. Our model is compared with various types of state-of-the-art general ReID methods, including one data augmentation-based method [23], three mask/attention-based methods [15, 18, 25], four striping-based methods [5, 7, 26, 46], two methods driven by some new loss functions [27, 29], and two global feature-based methods [13, 21], in addition to the seven occluded ReID methods in Table 3. The results on Market1501 and DukeMTMC are shown in Table 4. Our model achieves the best results in both R-1 and mAP on the two datasets among

Table 3 – R-1 and mAP of different methods on three occluded ReID datasets. The best performance is boldfaced.

Type	Method	OPReID		O-Duke		O-ReID	
		R-1	mAP	R-1	mAP	R-1	mAP
Occluded ReID	AdO [16]			44.5	33.2	-	-
	FD-GAN [9]	-	-	40.8	-	-	-
	DSR [12]	-	-	40.8	30.4	72.8	62.8
	PGFA [22]	48.1	48.3	51.4	37.3	-	-
	FPR [14]	48.7	46.6	-	-	78.3	68.0
	PVPM [8]	-	-	51.5	29.2	70.4	61.2
	HOReID [33]	47.6	46.0	55.1	43.8	**80.3**	70.2
General ReID	PB [27]	-	-	36.9	-	-	-
	PCB [26]	-	-	42.6	33.7		
	OSNet [46]		-	33.7	20.1	-	-
	ASB [21]	-	-	61.1	48.7	67.5	62.5
	Baseline [13]	61.1	62.1	62.8	51.9	68.2	63.4
	Ours	**65.8**	**67.2**	**69.0**	**57.2**	78.5	**72.9**

Table 4 – R-1 and mAP of different methods on two general ReID datasets. The best performance is boldfaced.

Type	Method	Market1501		DukeMTMC	
		R-1	mAP	R-1	mAP
Occluded ReID	AdO [16]	86.5	70.4	79.2	62.1
	FD-GAN [9]	90.5	77.7	80.0	64.5
	DSR [12]	83.6	64.3	-	-
	PGFA [22]	91.2	76.8	82.6	65.5
	FPR [14]	95.4	86.6	88.6	78.4
	PVPM [8]	93.1	82.3	84.9	71.8
	HOReID [33]	94.2	84.9	86.9	75.6
General ReID	PB [27]	91.7	79.6	84.4	69.3
	PCB [26]	92.3	77.4	81.9	65.3
	PN-GAN [23]	89.4	72.6	73.6	53.2
	HA-CNN [18]	91.2	75.7	80.5	63.8
	MGCAM [25]	83.8	74.3	80.7	66.4
	IANet [15]	94.4	83.1	87.1	73.4
	OSNet [46]	94.8	84.9	88.6	73.9
	BDB [7]	93.5	82.8	86.8	71.5
	ABD [5]	95.6	88.2	89.0	78.5
	ASB [21]	94.5	85.9	86.4	76.4
	Circle [29]	**96.1**	87.4	-	-
	Baseline [13]	95.4	88.6	89.9	79.8
	Ours	**96.1**	**89.3**	**91.1**	**81.3**

all general ReID and occluded ReID models.

When considering the performance across both occluded and general ReID datasets, HOReID [33] and Baseline [21] are two best competing models. Impressively, our model outperforms both of them on both occluded and general ReID datasets. This demonstrates excellent applicability of our model in real-world ReID applications with/without occlusions.

5.4. Ablation Study

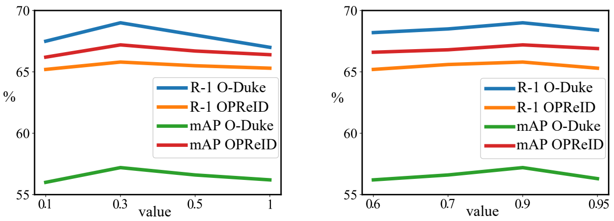
We evaluate the importance of five key components of our model, including Compound Batch Erasing (CBE), the BED loss, Disentangled Non-Local module (DNL), Reconstructive Pooling (RP) and Momentum-effect Elimination (ME). The results are provided in Table 5. From the results we can see that adding CBE and L_{bed} increase about 3-4 points in both R-1 and mAP on the occluded datasets. The two modules provide the main driving force for our model's superior performance. The modules, DNL and RP, enhance the backbone network and further increase the performance by 1-2 points. Note that the improvement driven by RP is applied to the occluded data only, since RP is designed specifically for feature reconstruction. Adding ME helps lift the performance by around 1 more point across both occluded and general ReID datasets, as the long-tailed problem generally exists in both types of data and ME helps alleviate the problem.

5.5. Parameter Sensitivity

The two key hyperparameters α and β, which respectively control the sensitivity of the proposed BED loss and the balance in the ME module, can be well tuned via cross validation. This section aims at providing some starting points of the parameter tuning based on our empirical results. Here $\alpha = 0.3$ and $\beta = 0.9$ are used by default and we vary one parameter with the other one fixed to examine its impact on the performance. The mAP and R-1 results are reported in Figure 5. Our model generally performs stably w.r.t. both parameters. In general, α needs to be sufficiently large. This is because when α is set to a very small value, the BED loss becomes non-aggressive, which can deteriorate the final performance. The performance is generally robust to the change of β.

Table 5 – R-1 and mAP of our model and its ablated variants.

CBE	BED	DNL	RP	ME	Occluded Dataset				General Dataset			
					O-Duke		OPReID		Market1501		DukeMTMC	
					R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
					62.8	51.9	61.1	62.1	95.4	88.6	89.9	79.8
✓					64.1	53.3	62.5	63.5	95.6	88.9	90.3	80.0
✓	✓				66.8	54.5	64.1	65.7	95.7	88.9	90.5	80.5
✓	✓	✓			67.0	54.8	64.8	65.9	95.9	89.0	90.7	80.8
✓	✓	✓	✓		68.1	56.1	65.2	66.4	96.0	89.0	90.8	80.9
✓	✓	✓	✓	✓	**69.0**	**57.2**	**65.8**	**67.2**	**96.1**	**89.3**	**91.1**	**81.3**



(a) Results vs. α (b) Results vs. β

Figure 5 – R-1 and mAP results w.r.t two parameters α and β.

6. Conclusions

This paper introduces a novel model to learn fine-grained discriminative features for occluded person ReID. This model has two key features: (i) it learns single-scale global features with single network backbone, which is significantly simpler than state-of-the-art occluded models that rely one or more auxiliary modules, but it can substantially outperform these contenders; and (ii) it achieves new SOTA performance on both occluded ReID and general ReID benchmarks, showing excellent applicability in real-life ReID applications. Through this model, we show for the first time that single-scale global-level features can outperform the popular multi-scale part-level features for occluded ReID, offering a new direction for exploring lightweight yet effective occluded ReID models. Further, we introduce a large-scale occluded ReID dataset that contains so far the most realistic occlusions and the largest number of identities and images, providing a significantly more faithful occluded ReID benchmark than existing datasets.

Acknowledgments YC and XB was in part supported by the NSFC project #61772057 and BNSF project #4202039.

References

[1] Honglong Cai, Zhiguan Wang, and Jinxing Cheng. Multi-scale body-part mask guided attention for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 1–8, 2019.

[2] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 0–0, 2019.

[3] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7291–7299, 2017.

[5] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *Int. Conf. Comput. Vis.*, pages 8351–8361, 2019.

[6] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1335–1344, 2016.

[7] Zuozhuo Dai, Mingqiang Chen, Siyu Zhu, and Ping Tan. Batch dropblock network for person re-identification and beyond. In *Int. Conf. Comput. Vis.*, 2019.

[8] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. Pose-guided visible part matching for occluded person reid. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11744–11752, 2020.

[9] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *Adv. Neural Inform. Process. Syst.*, pages 1222–1233, 2018.

[10] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETSW*, volume 3, pages 1–7, 2007.

[11] Luo Hao, Fan Xing, Zhang Chi, and Jiang Wei. Stnreid : Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification. *TMM*, 2020.

[12] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7073–7082, 2018.

[13] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631*, 2020.

[14] 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 *Int. Conf. Comput. Vis.*, pages 8450–8459, 2019.

[15] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-and-aggregation network for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9317–9326, 2019.

[16] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. Adversarially occluded samples for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5098–5107, 2018.

[17] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1–8, 2019.

[18] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2285–2294, 2018.

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Int. Conf. Comput. Vis.*, pages 2980–2988, 2017.

[20] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4099–4108, 2018.

[21] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 92–100, 2019.

[22] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *Int. Conf. Comput. Vis.*, pages 542–551, 2019.

[23] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In *Eur. Conf. Comput. Vis.*, pages 650–667, 2018.

[24] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(7):1655–1668, 2018.

[25] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1179–1188, 2018.

[26] Yifan Su, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling. In *Eur. Conf. Comput. Vis.*, pages 480–496, 2018.

[27] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *Eur. Conf. Comput. Vis.*, pages 402–419, 2018.

[28] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5693–5703, 2019.

[29] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6398–6407, 2020.

[30] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, and Shengjin Wang. Perceive where to focus: Learning visibility-aware part-level features for partial person reid. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 393–402, 2019.

[31] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *Adv. Neural Inform. Process. Syst.*, 2020.

[32] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Mancs: A multi-task attentional network with curriculum sampling for person reid. In *Eur. Conf. Comput. Vis.*, pages 365–381, 2018.

[33] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6449–6458, 2020.

[34] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM Int. Conf. Multimedia*, pages 274–282, 2018.

[35] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7794–7803, 2018.

[36] Cheng Yan, Guansong Pang, Xiao Bai, Changhong Liu, Ning Xin, Lin Gu, and Jun Zhou. Beyond triplet loss: person re-identification with fine-grained difference-aware pairwise loss. *IEEE Trans. Multimedia*, 2021.

[37] Hantao Yao, Shiliang Zhang, Richang Hong, Yongdong Zhang, Changsheng Xu, and Qi Tian. Deep representation learning with part loss for person re-identification. *IEEE Trans. Image Process.*, 28(6):2860–2871, 2019.

[38] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. Disentangled non-local neural networks. In *Eur. Conf. Comput. Vis.*, 2020.

[39] Yao Zhai, Xun Guo, Yan Lu, and Houqiang Li. In defense of the classification loss for person reid. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 50–58, 2019.

[40] 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.

[41] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1116–1124, 2015.

[42] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *Int. Conf. Comput. Vis.*, pages 4678–4686, 2015.

[43] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2138–2147, 2019.

[44] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Int. Conf. Comput. Vis.*, pages 3754–3762, 2017.

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

[46] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Int. Conf. Comput. Vis.*, pages 3702–3712, 2019.

[47] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. Occluded person re-identification. In *Int. Conf. Multimedia and Expo*, pages 1–6. IEEE, 2018.