Matching on Sets: Conquer Occluded Person Re-Identification Without Alignment

Mengxi Jia[1], Xinhua Cheng[2], Yunpeng Zhai[1], Shijian Lu[3], Siwei Ma[4, 5], Yonghong Tian[4, 5], Jian Zhang[1, 5*]

[1] School of Electronic and Computer Engineering, Peking University, China
[2] College of Computer Science, Sichuan University, China
[3] Nanyang Technological University, Singapore
[4] School of Electronics Engineering and Computer Science, Peking University, China
[5] Peng Cheng Laboratory, China
{mxjia, ypzhai, swma,yhtian, zhangjian.sz}@pku.edu.cn, 931736813@qq.com, shijian.lu@ntu.edu.sg

Abstract

Occluded person re-identification (re-ID) is a challenging task as different human parts may become invisible in cluttered scenes, making it hard to match person images of different identities. Most existing methods address this challenge by aligning spatial features of body parts according to semantic information (e.g. human poses) or feature similarities but this approach is complicated and sensitive to noises. This paper presents Matching on Sets (MoS), a novel method that positions occluded person re-ID as a set matching task without requiring spatial alignment. MoS encodes a person image by a pattern set as represented by a 'global vector' with each element capturing one specific visual pattern, and it introduces Jaccard distance as a metric to compute the distance between pattern sets and measure image similarity. To enable Jaccard distance over continuous real numbers, we employ minimization and maximization to approximate the operations of intersection and union, respectively. In addition, we design a Jaccard triplet loss that enhances the pattern discrimination and allows to embed set matching into deep neural networks for end-to-end training. In the inference stage, we introduce a conflict penalty mechanism that detects mutually exclusive patterns in the pattern union of image pairs and decreases their similarities accordingly. Extensive experiments over three widely used datasets (Market1501, DukeMTMC and Occluded-DukeMTMC) show that MoS achieves superior re-ID performance. Additionally, it is tolerant of occlusions and outperforms the state-of-the-art by large margins for Occluded-DukeMTMC.

Introduction

Person re-identification (re-ID) (Zheng, Yang, and Hauptmann 2016) aims to associate persons that non-overlapping cameras capture from different places and viewpoints. It is an indispensable component in video surveillance where persons are often the major objects of interest (Yang et al. 2019). Though we have observed impressive progress in holistic person re-ID in recent years (Wang et al. 2019b; Zhai et al. 2020a; Jia et al. 2020; Wang et al. 2020a), it is still a big challenge to match persons of different identities in cluttered scenes where different body parts often suffer from occlusions and become invisible. Occluded re-ID (Zhuo et al. 2018; Miao et al. 2019) remains a very open research challenge despite its great values in practical re-ID systems.

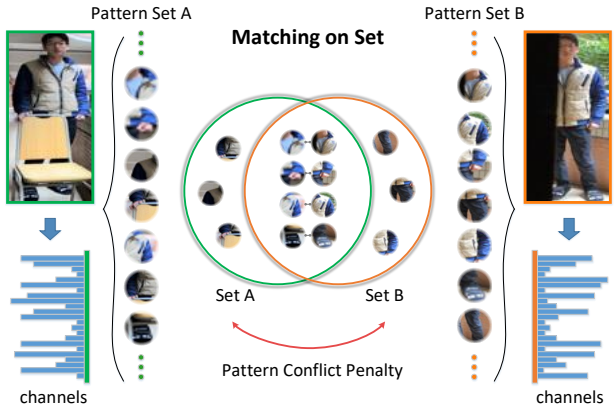

Figure 1: Illustration of Matching on Sets (MoS) for occluded person re-ID: MoS represents a person image by a 'global vector' that defines a pattern set. It matches occluded person images according to the Jaccard similarity between their pattern sets without requiring spatial alignment of person images. We design a Jaccard triplet loss for end-to-end network training and a conflict penalty mechanism for optimal inference.

The core challenge of occluded re-ID lies with two problems. The first problem comes from various occlusions that often cover up discriminative body parts. As a result, the trained model cannot extract sufficient representation information for re-ID. The second problem comes from the interference of obstacles that often share similar appearances as human bodies. Such obstacle interference deteriorates the extracted representation and makes the person matching complicated and prone to errors. Existing works address these challenges by adopting an intuitive approach that aligns the visible body parts and measures their similarities separately. The major line of research along this direction leverages external cues such as person masks (Song

*Corresponding author.

et al. 2018; He et al. 2019), semantic parsing (Kalayeh et al. 2018) or human poses (Miao et al. 2019; Gao et al. 2020; Wang et al. 2020b) to detect visible body parts for matching. On the other hand, the detection of visible parts is complicated and tends to fail to generate reliable visual cues under the presence of severe occlusions. Another line of research builds up part alignment relations according to the local feature similarity across person images (Zheng et al. 2015b; He et al. 2018a; Sun et al. 2019b). It is complicated as well and tends to produce mismatches because it often fails to differentiate human bodies from obstacles reliably.

In this paper, we present matching on sets (MoS), a novel network that treats occluded person re-ID as a set matching problem and accordingly bypasses the complicated and error-prone spatial alignment process. Based on the observation that a convolutional feature channel usually encodes a visual pattern (Zheng et al. 2017), MoS treats the convolutional feature vector of a person image as a pattern set and measures the person image similarity by measuring the similarity of the corresponding pattern sets as illustrated in Fig. 1. Specifically, we introduce Jaccard similarity coefficients as the metric to compute the similarity between pattern sets of person images. Since CNN feature vectors consist of real numbers, we introduce minimization and maximization to approximate the intersection and union operators in Jaccard similarity, which enables its computation on continuous real numbers and so differentiable in network training. We design a Jaccard triplet loss to incorporate the set matching operation into CNNs for end-to-end training. In addition, we design a conflict penalty mechanism that detects mutually exclusive patterns in the pattern union of image pairs and decreases their similarities for optimal inference.

The major contributions of this work can be summarized in three aspects. **First**, we formulate the occluded person re-ID as a set matching problem and design a network that measures the similarity between person images by using the Jaccard similarity coefficient between the corresponding pattern sets. **Second**, we design a novel Jaccard triplet loss that enables end-to-end network training with smoothing approximation. We also design a conflict penalty mechanism for the optimal inference that detects mutually exclusive patterns in the pattern union of image pairs and decreases their similarities accordingly. **Third**, our designed network is robust and does not require complicated and error-prone spatial alignment. It achieves superior re-ID performance under various occlusions yet without sacrificing performance over normal re-ID data with little occlusion.

Related Work

Person Re-identification

Person re-ID mainly focuses on the challenge which lies in the large intra-class and small inter-class variation caused by different viewpoints and poses, illumination conditions, camera configurations, etc. Existing methods can be broadly classified into two categories, one is committed to learning a discriminative feature representation for persons (Liao et al. 2015; Zheng et al. 2019; Wang et al. 2019a; Zhai et al. 2020b; Wang et al. 2020c), the other is learning similarity metrics to predict whether two images contain the same person (Zheng, Gong, and Xiang 2011; Sun et al. 2020; Jia et al. 2020). Most existing methods were developed for matching holistic person images that cannot tackle the occluded Re-ID problem. Specifically, when facing the occluded situations, those previous works mix information of the target person and obstacles into the final feature representation and usually fail in practical surveillance scenarios.

Occluded Person Re-ID

Existing works tackle Occluded re-ID by aligning the visible body regions and measures their similarities separately.

The major line of existing methods leverage external cues such as foreground segmentation or pose estimation to align the detected human bodies. (Miao et al. 2019) presents a pose guided feature alignment method to match the local patches of person images based on the human semantic key-points and use a pre-defined threshold of key-points confidence to determine the part visibility. (Gao et al. 2020) employs a pose-guided visible part matching algorithm (PVPM) that self-mine the part visibility via graph matching and adapt pose-guide attention accordingly. (Wang et al. 2020b) presents a framework utilizing key-points estimation to learn high-order relation information for discriminative features and human-topology information for robust alignment. By introducing extra semantic information as guidance, the above methods can accurately locate and align body parts; however, they inevitably cost much time to infer these external cues and may fail to generate reliable visual cues under the presence of severe occlusions.

Another line of research adopt a part-to-part matching strategy, which builds up part alignment relations according to the local feature similarity across person images. (Zheng et al. 2015b) presents a local patch-level matching model to capture the spatial layout information. (He et al. 2018a) reconstructs the feature map of a partial query sparsely from the feature map of holistic gallery images, and further improve it by a foreground-background mask to alleviate the contamination of occlusion in (He et al. 2019). (Sun et al. 2019b) presents a visibility-aware part model (VPM), which learns to perceive the visibility of regions through self-supervision. Recently, (Zhu et al. 2020) adopts clustering to learn the human semantic parsing to achieve the pixel-level alignment, which can locate both human body parts and personal belongings. These approaches flexibly match local features across person images in a self-guided way. On the other hand, they require complicated extra operations and often fail to differentiate human bodies from obstacles reliably, leading mismatch in the complex environment. Different from the above alignment-based approaches, our method addresses the occluded person re-ID by set matching on image pattern sets, which measures similarities between image pairs without spatial alignment.

Proposed Method

In this section, we first provide a novel perspective to formulate person re-ID as a pattern set matching problem and describe how to measure the similarity between person pattern sets using Jaccard similarity. Then we propose a Jaccard

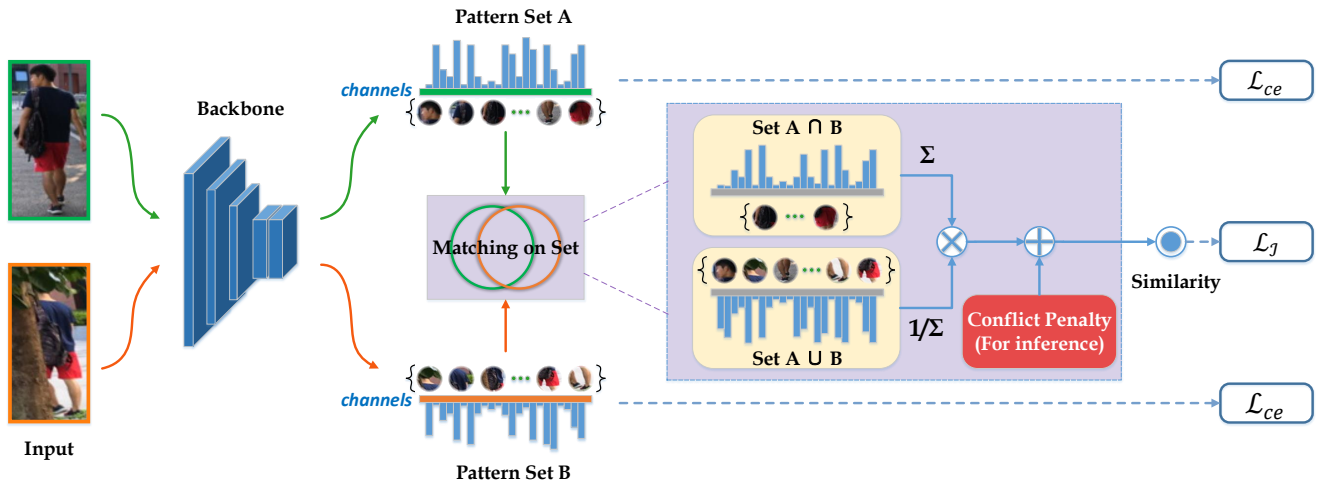

Figure 2: The framework of our proposed MoS: Given a pair of person images as *Input*, MoS first encodes them by two pattern sets (*Pattern Set A* and *Pattern Set B*) where each set is represented by a 'global vector' with each element capturing one specific visual pattern. It employs Jaccard similarity coefficient with a conflict penalty term as a metric to compute the distance between pattern sets and measure the image *Similarity*, more details to be described in Proposed Method

triplet loss with smoothing approximation that embeds the set matching task into networks for end-to-end training. Finally, we introduce a conflict penalty mechanism to improve the robustness of matching especially in occluded scenarios. An overview of our method is shown in Fig. 2.

Pattern Sets Matching

We view occluded re-ID as a pattern sets matching problem. For a person image $\mathbf{x}$, we first employ a CNN backbone to generate an image embedding $\mathbf{f} = \text{CNN}(\mathbf{x}) \in \mathbb{R}^C$, where C is the feature dimension. A non-linear activation function $\sigma(\cdot)$ is performed on $\mathbf{f}$ to obtain a vector $\mathbf{g} \in \mathbb{R}^C$, *i.e.,* $\mathbf{g} = \sigma(\mathbf{f})$. We consider $\mathbf{g}$ as a pattern set of $\mathbf{x}$, where each channel of $\mathbf{g}$ denotes a specific pattern such as glasses, collar, etc. A high value of a channel means that the corresponding pattern is visible in this image. Given two persons $\mathbf{x}_1$ and $\mathbf{x}_2$, we propose to measure their similarity by using the set similarity between their pattern sets $\mathbf{g}_1$ and $\mathbf{g}_2$:

$$\mathcal{S}(\mathbf{x}_1, \mathbf{x}_2) = \mathcal{S}_{set}(\mathbf{g}_1, \mathbf{g}_2), \tag{1}$$

where $\mathbf{g}_n = \sigma(\text{CNN}(\mathbf{x}_n))$ (n denotes the image index), $\mathcal{S}(\cdot,\cdot)$ denotes the similarity between images $\mathbf{x}_1$ and $\mathbf{x}_2$ for person re-ID, and $\mathcal{S}_{set}(\cdot,\cdot)$ is a kind of metric to measure similarities between sets.

Jaccard Similarity for Re-identification

The Jaccard Similarity is a similarity measure between two sets. It is defined as the size of the intersection divided by the size of the union of two sets. Given two sets $\mathbf{A}$ and $\mathbf{B}$, the Jaccard Similarity is computed using the following formula:

$$\mathcal{J}(\mathbf{A}, \mathbf{B}) = \frac{|\mathbf{A} \cap \mathbf{B}|}{|\mathbf{A} \cup \mathbf{B}|}, \tag{2}$$

where $|\cdot|$ denotes the cardinality of a set. Using C-dimension binary vectors $\{0,1\}^C$ to represent set $\mathbf{A}$ and $\mathbf{B}$, where each channel denotes a specific element, the Jaccard Similarity between these two sets is computed by:

$$\mathcal{J}(\mathbf{A}, \mathbf{B}) = \frac{\sum_{c=1}^{C} \mathbf{A}[c] \wedge \mathbf{B}[c]}{\sum_{c=1}^{C} \mathbf{A}[c] \vee \mathbf{B}[c]}, \tag{3}$$

where $\wedge$ and $\vee$ separately denote bit-wise AND and OR operators, and the operator $[\cdot]$ returns the element at position c in a vector or a set. For the re-ID problem, the similarities of person image pairs can be represented by the Jaccard similarities of their pattern sets. Considering that directly transforming embedding $\mathbf{f}$ to binary values may expand errors and reduce representation robustness, we use sigmoid function $\sigma(\cdot) = e^{\cdot}/(1+e^{\cdot})$ to activate the feature and obtain the pattern set $\mathbf{g} = \sigma(\mathbf{f}) \in \mathbb{R}^C$. To make the Jaccard Similarity adaptive to this continuous variable, we use minimization and maximization to approximate the bit-wise AND and OR operators in Eq. (3), respectively. For given pattern sets $\mathbf{g}_1$ and $\mathbf{g}_2$, the Jaccard Similarity is redefined by:

$$\mathcal{J}(\mathbf{g}_1, \mathbf{g}_2) = \frac{\sum_{c=1}^{C} min(\mathbf{g}_1[c], \mathbf{g}_2[c])}{\sum_{c=1}^{C} max(\mathbf{g}_1[c], \mathbf{g}_2[c])}, \tag{4}$$

where c denotes the channel index.

End-to-end Training

To embed the set matching into CNNs for end-to-end training, we propose a Jaccard triplet loss that uses Jaccard Distance instead of Euclidean Distance to measure dissimilarities of image pairs. To smooth the min/max operators, we introduce a Softmax-Jaccard Similarity:

$$\mathcal{J}_s(\mathbf{g}_1, \mathbf{g}_2) = \frac{\sum_{c=1}^{C} (\mathbf{w}_1^{min}[c] \cdot \mathbf{g}_1[c] + \mathbf{w}_2^{min}[c] \cdot \mathbf{g}_2[c])}{\sum_{c=1}^{C} (\mathbf{w}_1^{max}[c] \cdot \mathbf{g}_1[c] + \mathbf{w}_2^{max}[c] \cdot \mathbf{g}_2[c])}, \tag{5}$$

where $\mathbf{w}_k^{min}[c]$ and $\mathbf{w}_k^{max}[c]$ is softmin/softmax of $\mathbf{g}_k[c]$ along k:

$$\mathbf{w}_k^{min}[c] = \frac{e^{-\tau \cdot \mathbf{g}_k[c]}}{\sum_n e^{-\tau \cdot \mathbf{g}_n[c]}}, \quad \mathbf{w}_k^{max}[c] = \frac{e^{\tau \cdot \mathbf{g}_k[c]}}{\sum_n e^{\tau \cdot \mathbf{g}_n[c]}} \tag{6}$$

where $k = 1, ..., N$ (N is the number of training images) and τ is a smoothing factor. The Jaccard distance is defined by:

$$\mathcal{D}_{\mathcal{J}}(\mathbf{g}_1, \mathbf{g}_2) = 1 - \mathcal{J}_s(\mathbf{g}_1, \mathbf{g}_2), \tag{7}$$

The Jaccard triplet loss is defined by:

$$\mathcal{L}_{\mathcal{J}} = \sum_{n=1}^{N} [\delta + \mathcal{D}_{\mathcal{J}}(\mathbf{g}_n, \mathbf{g}_{n+}) - \mathcal{D}_{\mathcal{J}}(\mathbf{g}_n, \mathbf{g}_{n-})]_+, \tag{8}$$

where $\mathbf{x}_{n+}$ denotes the samples belonging to the same person with $\mathbf{x}_n$, $\mathbf{x}_{n-}$ denotes the samples belonging to different persons with $\mathbf{x}_n$. δ is a margin parameter. Considering the exponential transformation in sigmoid function may lead to gradient vanishing, we remove the sigmoid behind image embedding $\mathbf{f}$ during training. And instead we use Leaky ReLU on $\mathbf{f}$, a more linear activation function. In the next section, we validate that the final Jaccard triplet loss can enhance the pattern discrimination in the set vectors.

Additionally, we conduct a linear transformation after image embedding $\mathbf{f}$ and it outputs a predicted probability $\mathcal{P}_m(\mathbf{x}_n)$ of image $\mathbf{x}_n$ belonging to the identity m. The cross entropy loss with label smoothing is defined as:

$$\mathcal{L}_{ce} = -\sum_{n=1}^{N} \sum_{m=1}^{M} q_m \log \mathcal{P}_m(\mathbf{x}_n), \tag{9}$$

where $q_m = 1 - \xi + \frac{\xi}{M}$ if $m = y_n$, otherwise $q_m = \frac{\xi}{M}$. M is the class (person identity) number of the training set, and y_n is the ground-truth label of $\mathbf{x}_n$. ξ is a small constant, which is set as 0.1. The overall loss is therefore calculated as:

$$\mathcal{L} = \mathcal{L}_{\mathcal{J}} + \mathcal{L}_{ce}. \tag{10}$$

Pattern Conflict Penalty

We propose pattern conflict penalty during inference under the hypothesis that if there are incompatible patterns between two sets, the corresponding images are likely to belong to different persons. For instance, in general *blue shoes* and *black shoes* are a pair of incompatible patterns, and thus they won't show up together in the pattern sets of the same person. As shown in Fig. 3, we punish the similarity between two images according to the incompatible pattern pairs in their pattern union. The pattern conflict penalty mechanism can be divided into two steps: conflict mining and similarity penalty. For a pair of images, we define pattern concurrence matrix $\mathcal{M} \in \mathbb{R}^{C \times C}$ on their pattern union as:

$$\mathcal{M}(\mathbf{g}_1, \mathbf{g}_2) = (\mathbf{g}_1 \cup \mathbf{g}_2)(\mathbf{g}_1 \cup \mathbf{g}_2)^T, \tag{11}$$

where $\mathbf{g}_1 \cup \mathbf{g}_2 = max(\mathbf{g}_1, \mathbf{g}_2) \in \mathbb{R}^{C \times 1}$ denotes the pattern union of set $\mathbf{g}_1$ and $\mathbf{g}_2$. Each element in $\mathcal{M}$, denoted by $\mathcal{M}_{[i,j]}$, stands for whether pattern-i and pattern-j co-occur in the union of $\mathbf{g}_1$ and $\mathbf{g}_2$. For conflict mining, we compute

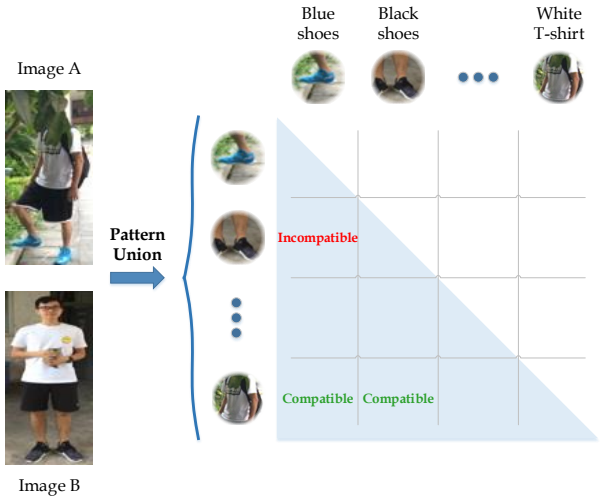

Figure 3: The proposed conflict penalty mechanism detects incompatible pattern pairs in the union of the pattern sets of two person images and lowers their similarity adaptively for optimal inference.

pattern concurrence matrices $\mathcal{M}$ between all positive sample pairs in the training set. Through bit-wise maximization, we obtain a priori max concurrence (PAC) matrix $\mathcal{M}^p$:

$$\begin{aligned} \mathcal{M}^p_{[i,j]} &= \max_{t,n} \mathcal{M}_{[i,j]}(\mathbf{g}_t, \mathbf{g}_n), \\ \text{s. t.} \quad & y_t = y_n, \\ & t, n = 1, ..., N. \end{aligned} \tag{12}$$

where $i, j = 1, ..., C$, y_t and y_n denote the ground-truth identities of image t and n, respectively. During inference, for given set pair $(\mathbf{g}_1, \mathbf{g}_2)$ we compare values in its pattern concurrence matrix $\mathcal{M}(\mathbf{g}_1, \mathbf{g}_2)$ with those in the priori max concurrence matrix $\mathcal{M}^p$, then punish their similarity if there are values in $\mathcal{M}(\mathbf{g}_1, \mathbf{g}_2)$ much higher than their corresponding ones in $\mathcal{M}^p$. The penalty term is calculated by:

$$\mathcal{CP}(\mathbf{g}_1, \mathbf{g}_2) = \sum_{i,j} max(0, e^{(\mathcal{M}_{[i,j]}(\mathbf{g}_1, \mathbf{g}_2) - \mathcal{M}^p_{[i,j]} - \varepsilon)} - 1), \tag{13}$$

where ε is an offset factor. The final similarity between $\mathbf{x}_1$ and $\mathbf{x}_2$ is computed by:

$$\mathcal{S}(\mathbf{x}_1, \mathbf{x}_2) = \mathcal{J}(\mathbf{g}_1, \mathbf{g}_2) - \lambda \mathcal{CP}(\mathbf{g}_1, \mathbf{g}_2), \tag{14}$$

where λ is a scale factor. **Algorithm 1** provides the overall description of our proposed MoS.

Experiments

Datasets and Evaluation Metrics

We evaluate MoS over two occluded re-ID datasets Occluded-DukeMTMC (Miao et al. 2019) and P-ETHZ (Zhuo et al. 2018). Occluded-DukeMTMC is a split of DukeMTMC-reID (Zheng, Zheng, and Yang 2017) which contains 15,618 training images, 17,661 gallery images, and 2,210 occluded query images. The experiments on this

Algorithm 1 Matching on Sets (MoS)

Input: Training/query/gallery set: $\mathcal{X}_{train}$, $\mathcal{X}_{query}$, $\mathcal{X}_{gallery}$
Output: Similarities $\mathcal{S}$

```
%Training stage
Initialize the CNN network parameters Θ.
for mini-batch B ⊂ X_train do
    Extract its pattern set g_n of each x_n in B using CNN.
    Calculate L_J by Eqs.(5, 6, 7, 8) and L_ce by Eq.(9).
    Optimize CNN parameters Θ according to Eq.(10).
end for
Calculate the PAC matrix M^p by Eq.(12).
%Inference stage
for each q ∈ X_query, g ∈ X_gallery do
    Calculate J(q, g) by Eq.(4) and CP(q, g) by Eq.(13).
    Calculate S(q, g) by Eq.(14).
end for
return S
```

dataset follow the standard setting (Miao et al. 2019). P-ETHZ is modified from ETHZ, which contains 3,897 images of 85 IDs. Following (Zhuo et al. 2018), we randomly select images of half identities for training and the rest for test. We also evaluate MoS over two widely used Holistic re-ID datasets Market-1501 (Zheng et al. 2015a) and DukeMTMC-reID (Zheng, Zheng, and Yang 2017) to test its generalizability.

For evaluations, we adopt Cumulative Matching Characteristic (CMC) curve and mean average precision (mAP) that are used widely in Re-ID evaluations.

Implementation Details

We adopt ResNet-50 (He et al. 2016) as the backbone network which is pre-trained over ImageNet (Krizhevsky, Sutskever, and Hinton 2017). For a fair comparison with methods whose backbone networks have a larger number of parameters, we employ another backbone that incorporates instance batch normalization (IBN) (Pan et al. 2018) into the ResNet50 for improving learning capacities without increasing computational costs. The models of the two networks are denoted by "MoS" and "MoS$_{w/ibn}$" (similarly hereinafter), respectively. During training, the input image is resized to 256×128 and augmented with random horizontal flipping, random erasing (Zhong et al. 2020) and random cropping. We warm up the model for 10 epochs with a linearly growing learning rate from 3.5×10^{-5} to 3.5×10^{-4}, and then decrease it by a factor of 0.1 at 40th and 70th epoch. The batch size is set to 64 and the Adam optimizer is adopted in model training. We fixed $\varepsilon = 0.1$ and $\lambda = 0.001$ in experiments, and implement the networks on PyTorch.

Comparison with the State-of-the-Art

We compare our method with state-of-the-art methods and Tables 1, 2 and 3 show the experimental results for both occluded and holistic person re-ID tasks. The compared methods can be classified into two categories according to the used backbones. One category is the mainstream which employs ResNet-50 directly or modifies ResNet-50 by introducing additional branches, attention subnets, dilated convolution, *etc*. The other category employs more powerful backbones with more parameters than ResNet50, such as ResNet101 or HRNet-W32. For fair comparison, we compare the two categories of methods with "MoS" and "MoS$_{w/ibn}$", respectively. The comparisons with the second category methods are largely for reference.

Methods	Rank-1	mAP
DIM (ArXiv 17)	21.5	14.4
Part Aligned (ICCV 17)	28.8	20.2
HACNN (CVPR 18)	34.4	26.0
Adver Occluded (CVPR 18)	44.5	32.2
PCB (ECCV 18)	42.6	33.7
Part Bilinear (ECCV 18)	36.9	-
FD-GAN (NIPS 18)	40.8	-
PGFA (ICCV 19)	51.4	37.3
HONet (CVPR 20)	55.1	43.8
DSR (CVPR 18)	40.8	30.4
SFR (ArXiv 18)	42.3	32.0
MoS *(Ours)*	**61.0**	**49.2**
ISP* (ECCV 20)	62.8	52.3
MoS$_{w/ibn}$ *(Ours)*	**66.6**	**55.1**

Table 1: Comparison over dataset Occluded-DukeMTMC: "*" highlight that different backbone is employed.

Methods	Rank-1	Rank-5	mAP
DGD (CVPR 16)	51.2	81.0	-
SVDNet (ICCV 17)	52.2	79.0	-
REDA (AAAI 20)	54.4	79.1	-
AFPB (ICME 18)	58.2	84.6	-
MoS *(Ours)*	76.2	84.8	64.0
MoS$_{w/ibn}$ *(Ours)*	**79.5**	**85.7**	**66.8**

Table 2: Comparison over dataset P-ETHZ

Results on Occluded-DukeMTMC Table 1 shows the comparison over the dataset Occluded-DukeMTMC. As Table 1 shows, three types of methods are compared including: **(1)** methods designed for holistic ReID (DIM (Yu et al. 2017), Part Aligned (Zhao et al. 2017), HACNN (Li, Zhu, and Gong 2018), Adver Occluded (Huang et al. 2018) and PCB (Sun et al. 2018)), **(2)** methods using extra pose information (Part Bilinear (Suh et al. 2018), FD-GAN (Ge et al. 2018), PGFA (Miao et al. 2019) and HONet (Wang et al. 2020b)), and **(3)** methods without using extra semantics (DSR (He et al. 2018a) and SFR (He et al. 2018b)). We can observe that MoS achieves 61.0% Rank-1 accuracy and 49.2% mAP and outperforms all three types of methods by large margins. For ISP that adopts a deeper backbone HRNet-W32, MoS$_{w/ibn}$ achieves higher Rank-1 and mAP and its parameter number is only 89% of ISP.

The MoS's superior performance is largely attributed to three aspects. First, the set matching is more robust than spatial alignment for re-ID in cluttered scenes. Second, the proposed Jaccard triplet loss helps learn more effective and discriminative pattern representations for the set matching inference. Third, the proposed pattern conflict penalty mecha-

Methods	Market-1501		DukeMTMC	
	Rank-1	mAP	Rank-1	mAP
IANet (CVPR 19)	94.4	83.1	87.1	73.4
MVPM (ICCV 19)	91.4	80.5	83.4	70.0
DMML (ICCV 19)	93.5	81.6	85.9	73.7
SFT (ICCV 19)	93.4	82.7	86.9	73.2
VCFL (ICCV 19)	89.3	74.5	-	-
Circle (CVPR 20)	94.2	84.9	-	-
PCB(ECCV 18)	92.3	77.4	81.8	66.1
PCB+RPP (ECCV 18)	93.8	81.6	83.3	69.2
AlignedReID(ArXiv18)	91.8	79.3	-	-
DSR (CVPR 18)	83.6	64.3	-	-
VPM (CVPR 19)	93.0	80.8	83.6	72.6
SPReID (CVPR 18)	92.5	81.3	-	-
MGCAM (CVPR 18)	83.8	74.3	46.7	46.0
Pose-transfer (CVPR18)	87.7	68.9	30.1	28.2
PSE (CVPR 18)	87.7	69.0	27.3	30.2
PGFA (ICCV 19)	91.2	76.8	82.6	65.5
AANet (CVPR 19)	93.9	82.5	86.4	72.6
HONet (CVPR 20)	94.2	84.9	86.9	75.6
MoS *(Ours)*	**94.7**	**86.8**	**88.7**	**77.0**
DCDS* (ICCV 19)	94.8	85.8	87.5	75.5
ISP* (ECCV 20)	95.3	88.6	89.6	80.0
MoS$_{w/ibn}$ *(Ours)*	**95.4**	**89.0**	**90.6**	**80.2**

Table 3: Comparison over datasets Market-1501 and DukeMTMC: The compared methods are grouped into four categories: global feature based, part feature based, external cues based and different backbone based.

nism improves the set similarity by utilizing compatible and incompatible inter-pattern relations comprehensively.

Results on P-ETHZ Table 2 shows the comparison of MoS with state-of-the-art methods on the dataset P-ETHZ. The compared methods include three holistic re-id methods (DGD (Xiao et al. 2016), SVDNet (Sun et al. 2017) and REDA (Zhong et al. 2020)), as well as an occluded re-id method AFPB (Zhuo et al. 2018). As Table 2 shows, MoS achieves the best performance under all three metrics and it outperforms the state-of-the-art by 18% in Rank-1 accuracy.

Results on Market-1501 and DukeMTMC-reID We compare MoS with a number of holistic re-ID methods to verify its effectiveness over the holistic re-id task. The compared methods include: (1) six methods using global features including IANet (Hou et al. 2019), MVPM (Sun et al. 2019a), DMML (Chen et al. 2019), SFT (Luo et al. 2019), VCFL (Liu and Zhang 2019) and Circle (Sun et al. 2020); (2) five methods using part features including PCB, PCB+RPP (Sun et al. 2018), AlignedReID (Zhang et al. 2017), DSR (He et al. 2018a) and VPM (Sun et al. 2019b); (3) eight methods using external cues including human-parsing based SPReID (Kalayeh et al. 2018) and MGCAM (Song et al. 2018); attribute information based AANet (Tay, Roy, and Yap 2019); human pose based Pose-transfer (Liu et al. 2018), PSE (Sarfraz et al. 2018), PGFA (Miao et al. 2019) and HONet (Wang et al. 2020b); (4) two methods using different backbone including DCDS with ResNet101 (Alemu, Shah, and Pelillo 2019) and ISP with HRNet-W32 (Zhu et al. 2020).

Table 3 shows experimental results. We can observe that MoS outperforms the 1^{st} category methods using global features consistently. In addition, MoS performs better than the 2^{nd} category methods using part features even when only global features are used, though part-based features are efficient for re-ID task (Sun et al. 2018). This shows that MoS can learn discriminative features without considering spatial distributions. Further, MoS outperforms the 3^{rd} category methods using external cues without leveraging any extra information. Last but not least, **MoS**$_{w/ibn}$ achieves competitive performance as compared with DCDS and ISP that employ ResNet101 and HRNet-W32 as backbone, respectively. This shows that MoS is a backbone-agnostic approach and can benefit from more powerful backbones.

Methods	**Metric**	Rank-1	mAP
Euclidean Triplet	Euclidean	60.2	49.5
Euclidean Triplet	Cosine	58.6	50.4
Euclidean Triplet	Jaccard	60.3	50.8
Jaccard Triplet	Jaccard	64.4	54.8
Jaccard Triplet	Jaccard + **CP**	66.6	55.1

Table 4: Ablation study over dataset Occluded-DukeMTMC. CP denotes our proposed conflict penalty mechanism.

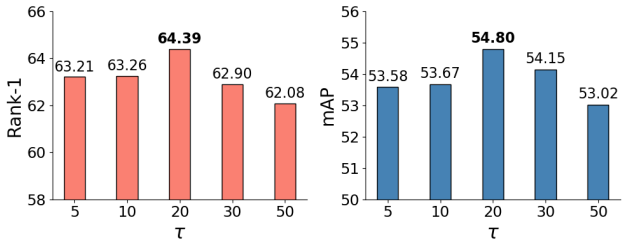

Figure 4: The smoothing factor τ in Eq. (6) affects Jaccard triplet loss and re-ID in mAP and Rank-1 accuracy.

Ablation Study

We conducted extensive ablation studies to evaluate each component of MoS. We used ResNet50 with IBN as backbone and performed ablation experiments over Occluded-DukeMTMC. Table 4 shows experimental results.

Perspective of Set Matching We first study the effect of positioning occluded re-ID as a set matching problem by adopting Jaccard similarity as metric in inference. In this experiment, we train a re-ID model with the original triplet loss with Euclidean distance, and compare its re-ID performance with models using different similarity metrics. As Table 4 shows, Jaccard similarity outperforms Euclidean and cosine similarity by small margins in mAP. This experiment shows that the very plain implementation of set matching can handle the occlusion challenge effectively. The benefits of employing set matching can be further demonstrated by introducing other relevant designs and operations.

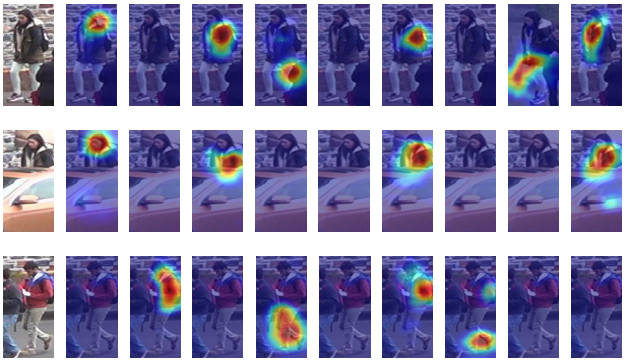

Figure 5: Visualization of global vectors and pattern sets: For the three sample images in the first column (the ones in the 1^{st} and 2^{nd} rows have the same identity, the last one has a distinct identity), the rest columns show the visual patterns of a number of randomly picked feature channels. The three visual patterns in each column have the same channel index.

Jaccard Triplet Loss We further study the effect of our designed Jaccard triplet loss as described in *End-to-end Training*. As Table 4 shows, employing Jaccard triplet loss improving the re-ID performance by large margins in both Rank-1 accuracy and mAP. The clear performance improvement shows that our designed Jaccard triplet loss guides to learn more discriminative features and patterns effectively while working with the set matching idea.

Pattern Conflict Penalty We conduct an extra experiment to study the effectiveness of our proposed conflict penalty (CP) mechanism and Table 4 shows experimental results. As Table 4 shows, the CP mechanism further improves re-ID performance, especially in Rank-1 accuracy. It helps as it penalizes the similarity of incompatible pattern pairs and improves the set similarity with better pattern matching.

The ablation studies show that the proposed set matching idea outperforms the *Baseline* (Triplet & Euclidean) by 6.4% in Rank-1 accuracy and 5.6% in mAP while working with the Jaccard triplet loss and conflict penalty mechanism. This demonstrates that the three components complement each other in achieving better occluded re-ID performance.

Discussion

Parameter Analysis The smoothing factor τ in Eq. (6) affects the Jaccard triplet loss and the occluded re-ID performance. We study this parameter by setting it to different values and checking the corresponding re-ID performance. Fig. 4 shows experimental results on Occluded-DukeMTMC dataset. Since a smaller τ usually leads to lager smoothing of both min and max values, it tends to obtain equal min and max values when τ is very small and degrades the re-ID accuracy. On the other hand, a large τ will sharpen the min and max values which usually leads to the difficulty of gradient propagation. Experiments show that MoS performs best when $\tau = 20$. From another perspective, MoS performs stably and is tolerant to the change of the parameter.

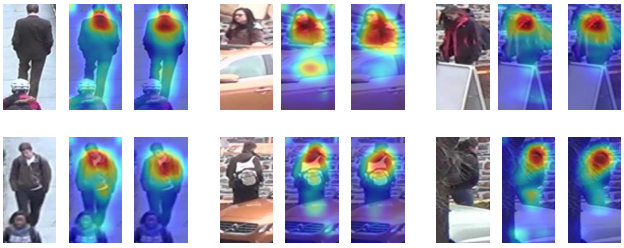

Figure 6: Illustration of feature activation maps: For each of the six groups of samples, the first image on the left is the original person image, the 2^{nd} and the 3^{rd} are the feature maps that are learned by the original triplet loss and our proposed Jaccard triplet loss, respectively.

Patterns Visualization We visualize several randomly picked channels of feature maps to demonstrate that a convolutional feature channel encodes a visual pattern effectively. As Fig. 5 shows, each channel learns a discriminative pattern in a local region. For the image of the same person (e.g. the 1^{st} and 2^{nd} person images), the same feature channel tends to locate similar local regions if not occluded. But for images of different persons (e.g. the 3^{rd} person), it tends to activate non-overlapping feature channels and produce different pattern sets.

We also visualize the activation of the whole feature map by averaging all feature channels. As Fig. 6 shows, our network can localize human body areas to learn discriminative representation even without explicitly modeling the localization of visible body parts. This nice property is largely attributed to the Jaccard triplet loss that guides the network to learn more discriminative body parts and patterns under the presence of noises and occlusions.

Conclusion

In this paper, we formulate the occluded person re-ID as a set matching problem and design a Matching on Sets (MoS) approach that measures the similarity between person images by using the Jaccard similarity coefficient between the corresponding pattern sets without spatial alignment. A Jaccard triplet loss is designed that enhances the pattern discrimination and allows to embed set matching into deep neural networks for end-to-end training. We also design a conflict penalty mechanism that decreases similarities of image pairs according to mutually exclusive patterns in their pattern union. Extensive experiments show that MoS achieves superior performance for occluded re-ID. We expect that the proposed MoS will inspire new insights and attract more interests for better occluded re-ID without spatial alignment.

Acknowledgments

This work was supported in part by Key-Area Research and Development Program of Guangdong Province (2019B121204008), National Natural Science Foundation of China (61902009) and Shenzhen Research Project (201806080921419290).

References

Alemu, L. T.; Shah, M.; and Pelillo, M. 2019. Deep Constrained Dominant Sets for Person Re-Identification. *International Conference on Computer Vision (ICCV)* 9854–9863.

Chen, G.; Zhang, T.; Lu, J.; and Zhou, J. 2019. Deep Meta Metric Learning. *International Conference on Computer Vision (ICCV)* 9546–9555.

Gao, S.; Wang, J.; Lu, H.; and Liu, Z. 2020. Pose-Guided Visible Part Matching for Occluded Person ReID. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 11741–11749.

Ge, Y.; Li, Z.; Zhao, H.; Yin, G.; Yi, S.; Wang, X.; et al. 2018. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *Advances in neural information processing systems*, 1222–1233.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

He, L.; Liang, J.; Li, H.; and Sun, Z. 2018a. Deep Spatial Feature Reconstruction for Partial Person Re-identification: Alignment-free Approach. *Conference on Computer Vision and Pattern Recognition (CVPR)* 7073–7082.

He, L.; Sun, Z.; Zhu, Y.; and Wang, Y. 2018b. Recognizing Partial Biometric Patterns. *ArXiv* abs/1810.07399.

He, L.; Wang, Y.; Liu, W.; Liao, X.; Zhao, H.; Sun, Z.; and Feng, J. 2019. Foreground-Aware Pyramid Reconstruction for Alignment-Free Occluded Person Re-Identification. *International Conference on Computer Vision (ICCV)* 8449–8458.

Hou, R.; Ma, B.; Chang, H.; Gu, X.; Shan, S.; and Chen, X. 2019. Interaction-And-Aggregation Network for Person Re-Identification. *Conference on Computer Vision and Pattern Recognition (CVPR)* 9309–9318.

Huang, H.; Li, D.; Zhang, Z.; Chen, X.; and Huang, K. 2018. Adversarially Occluded Samples for Person Re-identification. *Conference on Computer Vision and Pattern Recognition (CVPR)* 5098–5107.

Jia, M.; Zhai, Y.; Lu, S.; Ma, S.; and Zhang, J. 2020. A Similarity Inference Metric for RGB-Infrared Cross-Modality Person Re-identification. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 1026–1032.

Kalayeh, M.; Basaran, E.; Gokmen, M.; Kamasak, M.; and Shah, M. 2018. Human Semantic Parsing for Person Re-identification. *Conference on Computer Vision and Pattern Recognition (CVPR)* 1062–1071.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2017. Imagenet classification with deep convolutional neural networks. In *Communications of the ACM*, volume 60, 84–90.

Li, W.; Zhu, X.; and Gong, S. 2018. Harmonious Attention Network for Person Re-identification. *Conference on Computer Vision and Pattern Recognition (CVPR)* 2285–2294.

Liao, S.; Hu, Y.; Zhu, X.; and Li, S. Z. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2197–2206.

Liu, F.; and Zhang, L. 2019. View Confusion Feature Learning for Person Re-Identification. *International Conference on Computer Vision (ICCV)* 6638–6647.

Liu, J.; Ni, B.; Yan, Y.; Zhou, P.; Cheng, S.; and Hu, J. 2018. Pose Transferrable Person Re-identification. *Conference on Computer Vision and Pattern Recognition (CVPR)* 4099–4108.

Luo, C.; Chen, Y.; Wang, N.; and Zhang, Z. 2019. Spectral Feature Transformation for Person Re-Identification. *International Conference on Computer Vision (ICCV)* 4975–4984.

Miao, J.; Wu, Y.; Liu, P.; Ding, Y.; and Yang, Y. 2019. Pose-Guided Feature Alignment for Occluded Person Re-Identification. *International Conference on Computer Vision (ICCV)* 542–551.

Pan, X.; Luo, P.; Shi, J.; and Tang, X. 2018. Two at once: Enhancing learning and generalization capacities via ibn-net. In *European Conference on Computer Vision (ECCV)*, 464–479.

Sarfraz, M. S.; Schumann, A.; Eberle, A.; and Stiefelhagen, R. 2018. A Pose-Sensitive Embedding for Person Re-identification with Expanded Cross Neighborhood Re-ranking. *Conference on Computer Vision and Pattern Recognition (CVPR)* 420–429.

Song, C.; Huang, Y.; Ouyang, W.; and Wang, L. 2018. Mask-Guided Contrastive Attention Model for Person Re-identification. *Conference on Computer Vision and Pattern Recognition (CVPR)* 1179–1188.

Suh, Y.; Wang, J.; Tang, S.; Mei, T.; and Mu Lee, K. 2018. Part-aligned bilinear representations for person re-identification. In *European Conference on Computer Vision (ECCV)*, 402–419.

Sun, H.; Chen, Z.; Yan, S.; and Xu, L. 2019a. MVP Matching: A Maximum-Value Perfect Matching for Mining HardSamples, With Application to Person Re-Identification. In *International Conference on Computer Vision (ICCV)*, 6736–6746.

Sun, Y.; Cheng, C.; Zhang, Y.; Zhang, C.; Zheng, L.; Wang, Z.; and Wei, Y. 2020. Circle loss: A unified perspective of pair similarity optimization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 6398–6407.

Sun, Y.; Xu, Q.; Li, Y.; Zhang, C.; Li, Y.; Wang, S.; and Sun, J. 2019b. Perceive Where to Focus: Learning Visibility-Aware Part-Level Features for Partial Person Re-Identification. *Conference on Computer Vision and Pattern Recognition (CVPR)* 393–402.

Sun, Y.; Zheng, L.; Deng, W.; and Wang, S. 2017. Svdnet for pedestrian retrieval. In *International Conference on Computer Vision (ICCV)*, 3800–3808.

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *European Conference on Computer Vision (ECCV)*, 480–496.

Tay, C.-P.; Roy, S.; and Yap, K.-H. 2019. AANet: Attribute Attention Network for Person Re-Identifications. *Conference on Computer Vision and Pattern Recognition (CVPR)* 7127–7136.

Wang, G.; Gong, S.; Cheng, J.; and Hou, Z. 2020a. Faster Person Re-Identification. *European Conference on Computer Vision (ECCV)* .

Wang, G.; Yang, S.; Liu, H.; Wang, Z.; Yang, Y.; Wang, S.; Yu, G.; Zhou, E.; and Sun, J. 2020b. High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-Identification. *Conference on Computer Vision and Pattern Recognition (CVPR)* 6448–6457.

Wang, G.; Yang, Y.; Cheng, J.; Wang, J.; and Hou, Z. 2019a. Color-sensitive person re-identification. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Wang, G.; Zhang, T.; Cheng, J.; Liu, S.; Yang, Y.; and Hou, Z. 2019b. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *International Conference on Computer Vision (ECCV)*, 3623–3632.

Wang, G.-A.; Yang, T. Z.; Cheng, J.; Chang, J.; Liang, X.; Hou, Z.; et al. 2020c. Cross-Modality Paired-Images Generation for RGB-Infrared Person Re-Identification. *arXiv preprint arXiv:2002.04114* .

Xiao, T.; Li, H.; Ouyang, W.; and Wang, X. 2016. Learning deep feature representations with domain guided dropout for person re-identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 1249–1258.

Yang, F.; Yan, K.; Lu, S.; Jia, H.; Xie, X.; and Gao, W. 2019. Attention driven person re-identification. *Pattern Recognition* 86: 143–155.

Yu, Q.; Chang, X.; Song, Y.-Z.; Xiang, T.; and Hospedales, T. M. 2017. The Devil is in the Middle: Exploiting Mid-level Representations for Cross-Domain Instance Matching. *ArXiv* abs/1711.08106.

Zhai, Y.; Lu, S.; Ye, Q.; Shan, X.; Chen, J.; Ji, R.; and Tian, Y. 2020a. AD-Cluster: Augmented discriminative clustering for domain adaptive person re-identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 9021–9030.

Zhai, Y.; Ye, Q.; Lu, S.; Jia, M.; Ji, R.; and Tian, Y. 2020b. Multiple expert brainstorming for domain adaptive person re-identification. *European Conference on Computer Vision (ECCV)* .

Zhang, X.; Luo, H.; Fan, X.; Xiang, W.; Sun, Y.; Xiao, Q.; Jiang, W.; Zhang, C.; and Sun, J. 2017. AlignedReID: Surpassing Human-Level Performance in Person Re-Identification. *ArXiv* abs/1711.08184.

Zhao, L.; Li, X.; Zhuang, Y.; and Wang, J. 2017. Deeply-Learned Part-Aligned Representations for Person Re-identification. *International Conference on Computer Vision (ICCV)* 3239–3248.

Zheng, H.; Fu, J.; Mei, T.; and Luo, J. 2017. Learning multi-attention convolutional neural network for fine-grained image recognition. In *International Conference on Computer Vision (ICCV)*, 5209–5217.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015a. Scalable Person Re-identification: A Benchmark. In *International Conference on Computer Vision (ICCV)*, 1116–1124.

Zheng, L.; Yang, Y.; and Hauptmann, A. G. 2016. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984* .

Zheng, W.-S.; Gong, S.; and Xiang, T. 2011. Person re-identification by probabilistic relative distance comparison. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 649–656.

Zheng, W.-S.; Li, X.; Xiang, T.; Liao, S.; Lai, J.; and Gong, S. 2015b. Partial person re-identification. In *International Conference on Computer Vision (ICCV)*, 4678–4686.

Zheng, Z.; Yang, X.; Yu, Z.; Zheng, L.; Yang, Y.; and Kautz, J. 2019. Joint discriminative and generative learning for person re-identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2138–2147.

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in Vitro. In *International Conference on Computer Vision (ICCV)*, 3774–3782.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random Erasing Data Augmentation. In *International Joint Conference on Artificial Intelligence(AAAI)*, 13001–13008.

Zhu, K.; Guo, H.; Liu, Z.; Tang, M.; and Wang, J. 2020. Identity-Guided Human Semantic Parsing for Person Re-Identification. *ArXiv* abs/2007.13467.

Zhuo, J.; Chen, Z.; Lai, J.; and Wang, G. 2018. Occluded Person Re-Identification. In *International Conference on Multimedia and Expo (ICME)*, 1–6.