Semi-attention Partition for Occluded Person Re-identification

Mengxi Jia[1, 3*], **Yifan Sun**[3], **Yunpeng Zhai**[4], **Xinhua Cheng**[4], **Yi Yang**[5], **Ying Li**[2†]
[1] School of Software and Microelectronic, Peking University, Beijing, China
[2] National Engineering Center of Software Engineering, Peking University, Beijing, China
[3] Baidu Research
[4] Peking University, China
[5] College of Computer Science and Technology, Zhejiang University, China
{mxjia, ypzhai, li.ying}@pku.edu.cn,

Abstract

This paper proposes a Semi-Attention Partition (SAP) method to learn well-aligned part features for occluded person re-identification (re-ID). Currently, the mainstream methods employ either external semantic partition or attention-based partition, and the latter manner is usually better than the former one. Under this background, this paper explores a potential that the "weak" semantic partition can be a good teacher for the "strong" attention-based partition. In other words, the attention-based student can substantially surpass its noisy semantic-based teacher, contradicting the common sense that the student usually achieves inferior (or comparable) accuracy. A key to this effect is: the proposed SAP encourages the attention-based partition of the (transformer) student to be *partially* consistent with the semantic-based teacher partition through knowledge distillation, yielding the so-called semi-attention. Such partial consistency allows the student to have both consistency and reasonable conflict with the noisy teacher. More specifically, on the one hand, the attention is guided by the semantic partition from the teacher. On the other hand, the attention mechanism itself still has some degree of freedom to comply with the inherent similarity between different patches, thus gaining resistance against noisy supervision. Moreover, we integrate a battery of well-engineered designs into SAP to reinforce their cooperation (*e.g.*, multiple forms of teacher-student consistency), as well as to promote reasonable conflict (*e.g.*, mutual absorbing partition refinement and a supervision signal dropout strategy). Experimental results confirm that the transformer student achieves substantial improvement after this semi-attention learning scheme, and produces new state-of-the-art accuracy on several standard re-ID benchmarks.

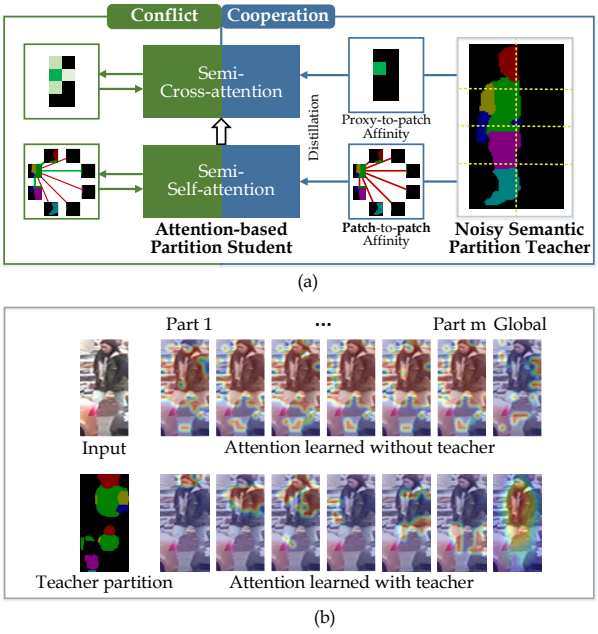

Figure 1: (a) Semi-attention partition is learned by cooperation (patch/proxy-to-patch consistency) and reasonable conflict with the teacher. (b) In SAP, the student (with teacher guidance) learns more discriminative part attention than the pure-attention baseline, and further surpasses the noisy teacher through reasonable student-teacher conflict

Introduction

This paper considers the occluded person re-identification (re-ID) (Zheng, Yang, and Hauptmann 2016; Zhai et al. 2020a). Basically, re-ID aims to identify the same person across multiple non-overlapping camera views. Occluded re-ID further raises a critical challenge, *i.e.*, the persons are partially occluded during observation. This challenge is very common among realistic re-ID scenarios, where the pedestrian sometimes appears in cluttered or crowded scenes. The occlusion problem brings additional distractions from the background, therefore significantly compromising the re-ID accuracy.

Learning well-aligned part features is a mainstream solution for occluded re-ID. It conducts part-to-part comparison (He et al. 2018b,a, 2019) by aligning the visible body parts and measuring their similarities separately. According to the image partition strategy, existing occluded re-ID methods can be divided into two categories: one approach (Gao et al. 2020; Wang et al. 2020; He and Liu 2020) uses external cues from semantic partition (*e.g.*, pose estimation and human parsing), while the other one (Sun et al. 2019a; Li et al. 2021; Zhu et al. 2021; He et al. 2021) uses attention mechanism for partition. Currently, the semantic-based approach is inferior

*Work done during an internship at Baidu Research.
†Corresponding author.

to the attention-based approach, because the off-the-shelf semantic partition is very inaccurate on re-ID images.

This paper explores a novel potential that the inaccurate semantic partition can be a good teacher for the attention-based partition. In other words, we intend to use an attention-based student to learn and gain clear benefit from the noisy semantic-based teacher. Our motivation indeed contradicts the common situation in teacher-student distillation framework, *i.e.*, the teacher is usually more accurate than the student (Bagherinezhad et al. 2018; Yang et al. 2019). More differences between our motivation and the popular student-teacher distillation are discussed in related work.

To facilitate this potential, we propose a novel Semi-Attention Partition (SAP) method for occluded re-ID. As illustrated in Fig. 1 (a), SAP has three components, *i.e.*, 1) a (noisy) semantic partition teacher, 2) an attention-based partition student, and 3) a semi-attention teaching strategy. Given an input pedestrian image, the student learns several part features under the cooperation of the self-attention mechanism and the semantic teacher supervision, yielding the so-called semi-attention. Since the attention mechanism is necessary for the student, SAP uses the transformer backbone as a natural choice. Specifically, the teacher supervision is enforced with two consistency constraints, *i.e.*, patch-to-patch consistency and patch-to-proxy consistency:

• *Patch-to-patch consistency.* If the teacher considers two patches as belonging to the same (different) semantic part(s), the student should make them have strong (weak) mutual attention.

• *Proxy-to-patch consistency.* If the teacher considers a set of patches potentially belongs to a specific part, the student should make the corresponding proxy have strong attention towards them.

An important reason making the student gain benefit from its noisy teacher is: the student is likely to have both cooperation and reasonable conflict with the teacher. On the one hand, the attention is guided by the semantic partition. On the other hand, the attention mechanism itself still has some degree of freedom to follow the inherent similarity between different patches, gaining resistance against some noisy supervision. Apart from the two forms of consistency for reinforcing the teacher-student cooperation, we also introduce a battery of well-engineered designs to promote their reasonable conflict, *e.g.*, SAP involves a mutual absorbing strategy to further refine partition, and performs dropout on the pixel-wise supervision signals from the teacher to effectively reduce the noises. Consequently, the cooperation and reasonable conflict jointly facilitate more stable and accurate partition than the pure attention and the semantic partition teacher, therefore bringing substantial improvement.

SAP has an advantage during testing: once the transformer student finishes its course from the noisy semantic teacher, it no longer needs the guidance from the teacher and is independently competent for extracting features. This advantage distinguishes our SAP from some part-refining methods which first conduct coarse partition and then refine the raw parts. In experiments, the visualization in Fig. 1 (b) shows that the learned parts of SAP are reasonably accurate, and the achieved re-ID accuracy further confirms that SAP improves occluded re-ID. For example, on Occluded-Duke, SAP achieves an mAP of 62.2% and a rank-1 accuracy of 70.0%, setting a new state-of-the-art.

In summary, our contribution is as follows: (i) We propose a novel teacher-student strategy called "Semi-Attention Partition" which explores a potential that the "bad" semantic partition can be a good teacher for the attention-based partition. (ii) We introduce some well-engineered designs to encourage the student to have both cooperation and reasonable conflict with noisy teacher. We find such cooperation and conflict are the important reasons that the student gains benefit from noisy supervision. (iii) We conduct extensive experiments on both occluded re-ID and holistic Re-ID benchmark datasets, demonstrating that our approach performs favorably against state-of-the-art methods.

Related Work

Occluded Person Re-ID

The difficulty of occluded re-ID mainly lies in the lack of discriminative characters caused by the invisible body parts and the disturbance of the obstruction regions. To tackle these problems, multi-part based representations were proposed to capture accessible characters of visible human parts and then assemble the part features to form the final descriptor.

Along this direction, one major line of works exploit attention to automatically localize discriminative human parts only with identity labels. Early methods embed the attention mechanism in the convolutional neural network (CNN) that predicted the human parts (Sun et al. 2019a; Li et al. 2021) or refined heuristic part division with learned within-part consistency (Sun et al. 2019b). However, the small receptive field of CNN and the single-head attention structure fail to capture the full image context information and easily ignore fine-grained characteristic under a large amount of occlusion noise. Vision transformer based model like ViT (Dosovitskiy et al. 2020) has achieved better results in person re-ID (Li et al. 2021; Ma, Zhao, and Li 2021; Jia et al. 2022; Zhu et al. 2021; He et al. 2021) by taking as core multi-head self-attention mechanism (Vaswani et al. 2017), thanks to the capability to capture long-range correlation of long sequences. However, learning such discriminative attentions requires pre-training with hundreds of millions of images on the classification task due to the lack of inductive bias and the larger amount of parameters (Touvron et al. 2021). And when adapting to the downstream task like occluded re-ID, it fails to learn effective attention end-to-end automatically under the limitation of the insufficient training data.

The other line of part-based works leverage external cues from the semantic partition with the assistance of pose estimation or human parsing. Miao (Miao et al. 2019a) proposes a pose-guided feature alignment method (PGFA), taking advantage of the human semantic key-points to guide the matching of probe and gallery images. Gao (Gao et al. 2020) presents a pose-guided visible part matching algorithm (PVPM) which jointly learns features and predicts the part visibility with attention heatmaps guided by pose estimation and graph matching accordingly. Wang (Wang et al. 2020) proposes a framework jointly modeling high-order relation and human-

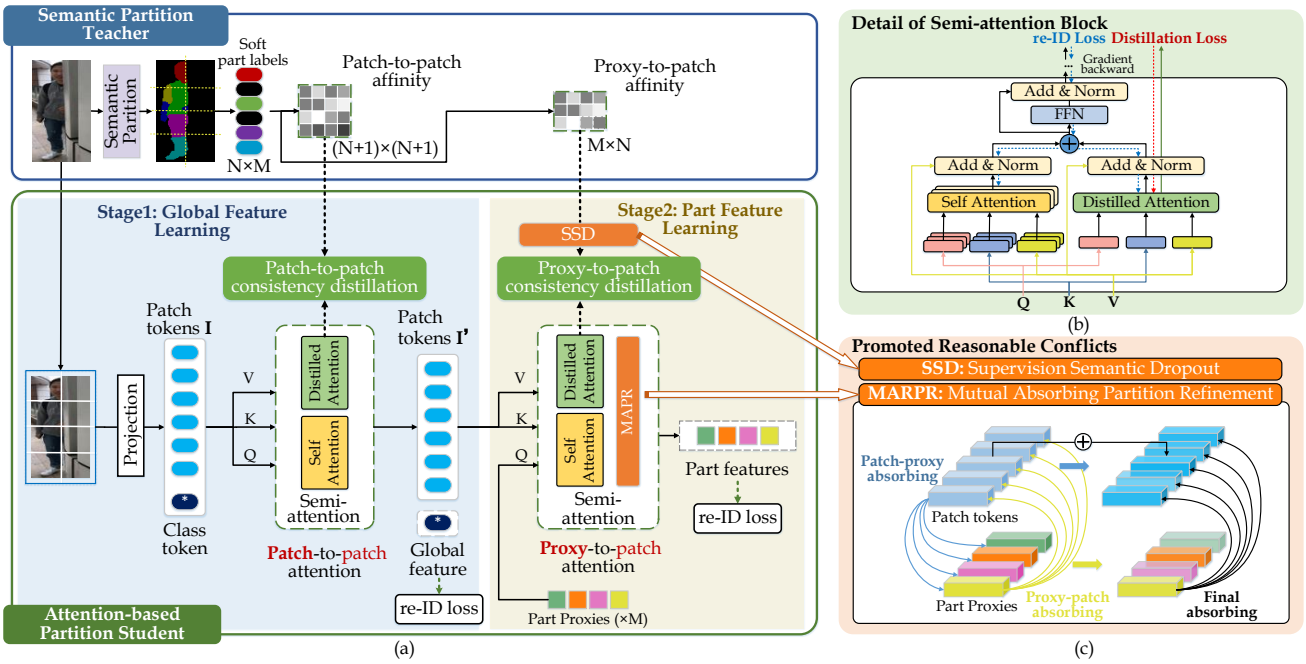

Figure 2: (a) The pipeline of SAP. It consists of a semantic partition teacher and a transformer-based student. The student has two feature learning stages, *i.e.*, the global feature learning and the part feature learning. In these two stages, the student respectively uses patch-to-patch attention and proxy-to-patch attention to refine and to collect part features. These two attention modules both combine self attention and distilled attention (which is supervised by the teacher model), as detailed in (b). Since the teacher is noisy, we insert two modules to encourage the student to have reasonable conflict with the noisy teacher, *i.e.*, the supervision semantic dropout (SSD) and the mutual absorbing partition refinement (MAPR), as shown in (c).

topology information by utilizing key-points estimation for robustly aligned features. Recently, some works adopt transformer architecture and incorporate off-the-shelf external cues (Wang et al. 2022a; Cheng et al. 2022). However, the inference of extra modules costs more time inevitably, and the generated semantic labels are untrustworthy when utilized in in an off-the-shelf manner. In SAP, the distillation procedure avoids direct exposure to external partition noises, and the reasonable teacher-student conflict further suppresses the noises.

Knowledge Distillation

Distilling knowledge (Hinton et al. 2015) from well-trained neural networks and transferring it to another model/network has been widely studied in recent years (Yim et al. 2017; Anil et al. 2018). Recent distillation process can be roughly categorized into three genres, *i.e.*, teacher-student framework (Lopez-Paz et al. 2015; Touvron et al. 2021), peer-teaching framework (Zhang et al. 2018; Zhai et al. 2020b, 2022) or self-distillation framework (Yun et al. 2020). Our method is most relevant to the teacher-student framework which uses the soft output (class probabilities or feature representations) of a teacher network to supervise a student network so as to make student models learn discrimination ability from teacher models. Normally, teacher-student knowledge transfer is performed from an over-parameterized teacher model to a lightweight student model for parameter reduction. And the student model is usually trained with data in the same domain as the teacher to mimic the output of it and inherit the dark knowledge. Our work considers a new situation where the teacher model is not the expert in the domain of the student. How can a stronger learner inherit the useful knowledge from a noisy teacher? The difference between our approach and typical teacher-student framework is summarized to three aspects: 1) The teacher and student models are trained by different data for different tasks. 2) It allows a weaker teacher than the student model in architecture or accuracy. 3) It distills different knowledge of inductive bias such as attention relation instead of classification logits.

The Proposed Approach

Overview

Our approach consists of a semantic partition teacher and an attention-based partition student, as illustrated in Fig. 2, and the goal is to learn the student from the noisy teacher and surpass it on occluded re-ID. The **teacher** $\mathcal{T}$ is a semantic partition model which has been pretrained on other datasets. Given a training image for occluded re-ID, the teacher model partitions it into multiple semantic parts for supervising the attention of the transformer student.

The **student** $\mathcal{S}$ uses a stack of transformer layers as its backbone, *i.e.*, **1)** a global feature learning stage featured for patch-to-patch attention and **2)** a sub-sequential part feature learning stage featured for proxy-to-patch attention, as

illustrated in Fig. 2.

• The first stage learns global feature representations using the class token. Specifically, the class token represents the global feature and is supervised through popular re-ID loss functions (*i.e.*, the cross-entropy loss and the triplet loss). Although the first stage learns only the global feature, we introduce patch-to-patch attention among all the patch tokens. We think this attention can make the patches belonging to a same part become more similar, therefore benefiting the sub-sequential part feature learning stage.

• The second stage learns part-level features by collecting different patches into several part features. The collection is accomplished with a proxy-to-patch attention. Specifically, the student inserts multiple learnable part proxies and constructs proxy-to-patch attention. Each part proxy uses proxy-to-part attention to absorb part-specific information from the patch tokens and generates a corresponding part-level feature. All the part-level features are then concatenated into the feature representation, which is supervised through popular re-ID loss functions.

For testing, we concatenate the global and part features to get the final representation.

In the following section, we first elaborate the global feature learning stage and the part feature learning stage, with emphasis on their novel attention modules (*i.e., the patch-to-patch attention and the proxy-to-patch attention*), respectively. Importantly, both the patch-to-patch and proxy-to-patch attention consist of a canonical self-attention using multi-head structures and a distilled attention (which is supervised by the teacher). We explore an effective structure to integrate the self-attention and the distilled attention into a semi-attention block, as shown in Fig.2. More details and analysis are provided in the appendix. Since the semi-attention need the supervision from the noisy teacher, we then introduce how to promote reasonable conflict between the student attention and the noisy teacher supervision.

Global Feature Learning

Given an input person image, we first follow ViT (Dosovitskiy et al. 2020) to transform it into multiple patch tokens through a linear projection. Based on these patch tokens, the **student** $\mathcal{S}$ learns patch-to-patch attention to facilitate information propagation among inherently-similar patches. We note that the patch-to-patch attention is indeed a basic module in ViT (Dosovitskiy et al. 2020). However, our patch-to-patch attention is different because it not only contains the standard self attention which is implemented with a Multi-head Self-Attention (MSA), but also contains another distilled attention (DA) supervised by the noisy teacher.

Specifically, for the **student** $\mathcal{S}$, suppose an person image $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ is given, where H, W and C denote the height, width and the channel number of image, respectively. We first split $\mathbf{x}$ into a sequence of N flattened image patches $\mathbf{x}_n$ $(n = 1, 2, ..., N)$, and employ a trainable linear projection $\mathcal{F}_{pe}(\cdot)$ to map the patches to D dimension patch tokens. The patch size is P with s overlapping pixels, and $N = \frac{H-S}{P-S} \times \frac{W-S}{P-S}$ is the number of patches. Moreover, a extra learnable class token $\mathbf{x}_{glo} \in \mathbb{R}^D$ is added to represent the global information. During first stage, the input patch token sequences as follows:

$$\mathbf{I} = [\mathbf{x}_{glo}; \mathcal{F}_{pe}(\mathbf{x}_1); \mathcal{F}_{pe}(\mathbf{x}_2); ...; \mathcal{F}_{pe}(\mathbf{x}_N)] + \mathbf{E}_{pos}, \quad (1)$$

where $\mathbf{E}_{pos} \in \mathbb{R}^{(N+1)\times D}$ is the position embeddings to retain positional information.

The global feature is represented with the final output of the class token $\mathbf{x}_{glo} \in \mathbb{R}^D$. Following the common practice in re-ID, we use a classification loss (the cross-entropy loss) and a pairwise loss (the softmax triplet loss) for supervising the global feature learning.

While the global feature learning procedure is arguably standard, our novelty lies in that we use a semi-supervised patch-to-patch attention to refine the part features (although the global feature learning stage does not supervise the part features).

Semi-supervised patch-to-patch attention. In parallel to the **student** $\mathcal{S}$, an off-the-shelf partition model, *i.e.*, the **teacher** $\mathcal{T}$ obtains a partition map $\mathbf{m} \in \{0, 1, ..., M-1\}^{H\times W}$ from the input image. Each element in the map denotes the part label of the corresponding pixel and M is the total number of semantic parts. Among M semantic parts, the $\mathcal{B}$-th part denotes the background. Similar with student model, we split $\mathbf{m}$ into a sequence of N patches $\{\mathbf{m}_n | n = 1, 2, ..., N\}$ with patch size $P \times P$, and encode the labels to one-hot forms by $\mathcal{G}_{oh}(\mathbf{m}) \in \{0, 1\}^{N\times P\times P\times M}$. Considering a single patch may contain pixels from different semantic parts, we assigned a soft label to each patch for robustness. Specifically, we first get the frequency of each semantic parts by averaging the $\mathcal{G}_{oh}(\mathbf{m})$ over both the width and height dimensions, and then perform the scaled softmax function to probabilize the label. The final part labels $\hat{\mathbf{m}}_p$ for patch sequences is defined as

$$\hat{\mathbf{m}} = softmax(\lambda \sum_{i=1}^{P}\sum_{j=1}^{P} \mathcal{G}_{oh}(\mathbf{m})_{[i,j]}) \quad (2)$$

where $\mathcal{G}_{oh}(\cdot)$ denotes the one-hot encoding function, $[i, j]$ is spatial position along width and height, and λ is a hyper-parameter for sharping the probability distribution ($\lambda \gg 1$).

For patch-to-patch attention, the DA begins by forming a set of queries $\mathbf{I}W_q$, and key-value pairs $\mathbf{I}W_k$, $\mathbf{I}W_v$. The $W_{q,k,v}$ are the learnable matrices for input patch tokens $\mathbf{I}$, and the original dot-product attention is defined as

$$\text{DA}_{patch-patch}(\mathbf{I}) = softmax(\frac{\mathbf{I}W_q(\mathbf{I}W_k)^T}{\sqrt{D}})\mathbf{I}W_v, \quad (3)$$

where $\mathbf{I}W_q(\mathbf{I}W_k)^T \in \mathbb{R}^{(N+1)\times(N+1)}$ is regarded as the patch-to-patch affinity scores. To learn the semantic relations across image patches, the distilling target is defined by the similarity of part labels among patches. Specifically, it is computed by the self-product of the label sequence $\hat{\mathbf{m}}\hat{\mathbf{m}}^T \in \mathbb{R}^{N\times N}$. Since the class token represents the global feature, the target for it is the person foreground (any visible human parts). Then the semantic similarity targets $\mathbf{A}_{patch} \in \mathbb{R}^{(N+1)\times(N+1)}$ for distilling is defined by

$$\mathbf{A}_{patch} = \begin{bmatrix} 1 & \mathbf{m}_f^T \\ \mathbf{m}_f & \hat{\mathbf{m}}\hat{\mathbf{m}}^T \end{bmatrix}, \quad \mathbf{m}_f = \mathbb{I}_{\hat{\mathbf{m}}[\mathcal{B}]<\gamma} \in \{0,1\}^{N\times 1} \tag{4}$$

where $\mathbf{m}_f$ denotes whether the patch refers to the foreground. $\mathcal{B}$ is the part index of background and $\hat{\mathbf{m}}[\mathcal{B}] \in \mathbb{R}^{N\times 1}$ denotes the probability of each patch belonging to background. γ is a threshold. Therefore, the distillation loss for patch-to-patch consistency is defined by the cross-entropy between the affinity scores and the semantic similarity targets:

$$\mathcal{L}_{patch} = \alpha \sum softmax(\mathbf{A}_{patch}) \log(softmax(\frac{\mathbf{I}W_q(\mathbf{I}W_k)^T}{\sqrt{D}})), \tag{5}$$

where α is the weight the distillation loss.

Part Feature Learning

We recall that in the global feature learning stage, the student outputs a sequence of patch features $\mathbf{I}^{'}$ for patch tokens and a global feature $\mathbf{f}_{glo}$ corresponding to the class token $\mathbf{x}_{glo}$. In the following part feature learning stage, the student use proxy-to-patch attention to collect the information scattered among different patches into several part features. To this end, the student model use a set of learnable part proxies to represent the part features. Each proxy absorbs patch features $\mathbf{I}^{'}$ into itself through multiple layers of proxy-to-patch attention. The proxy-to-patch attention here is also implemented by semi attention block with MSA and DA, as illustrated in Fig. 2.

Semi-supervised proxy-to-patch attention. Specifically, given a set of learnable part proxies $\mathbf{Q}_{proxy} \in \mathbb{R}^{M\times d}$, the distilled attention for proxy-to-patch attention contains queries $\mathbf{Q}_{proxy}$ and key-values pairs $\mathbf{I}^{'}W_k^{'}$, $\mathbf{I}^{'}W_v^{'}$. The $W_{k,v}^{'} \in \mathbb{R}^{D\times d}$ map patch features $\mathbf{I}^{'}$ to d dimension. Then the DA here is defined as

$$\text{DA}_{proxy-patch}(\mathbf{Q}_{proxy}, \mathbf{I}^{'}) = \text{softmax}(\frac{\mathbf{Q}_{proxy}(\mathbf{I}^{'}W_k^{'})^T}{\sqrt{d}})\mathbf{I}^{'}W_v^{'}, \tag{6}$$

where $\mathbf{Q}_{proxy}(\mathbf{I}^{'}W_k^{'})^T \in \mathbb{R}^{M\times N}$ is regarded as the proxy-to-patch affinity scores. To learn the consistency of it from teacher, the distilling targets is defined by the transposition of the part labels $\hat{\mathbf{m}}_p$ for patch token sequence, representing which patches are responding to the proxy. And the distillation loss for the proxy-to-patch consistency is defined as

$$\mathcal{L}_{proxy} = \alpha \sum softmax(\hat{\mathbf{m}}^T) \log(\text{softmax}(\frac{\mathbf{Q}_{proxy}(\mathbf{I}^{'}W_k^{'})^T}{\sqrt{d}})). \tag{7}$$

Through part feature learning, each proxy absorbs corresponding part-level information from the patches and becomes an individual part feature. Given the final state of these proxies, we concatenate all of them except the background proxy as the part-level representation. The concatenated part-level representation is supervised with the popular re-ID losses (softmax + triplet loss), which are same as the supervision for the global feature.

Good practices to promote reasonable conflict Although the prior information of semantic partition from the teacher is well utilized by cooperation, the student is hard to surpass the teacher due to the noisy supervision from the domain gap of data. To this end, we further promote reasonable conflict to gain resistance against the noisy supervision of the teacher.

1) Mutual Absorbing Partition Refinement. To surpass the noisy teacher supervisions, we further introduce an Mutual Absorbing Partition Refinement strategy to learn reasonable conflicts between proxies and patches for adaptively refining each other, as illustrated in Fig. 2. Specifically, if a patch token potentially belongs to a specific part, this token first absorbs information from the corresponding part proxy (i.e. the left yellow box) by a reverse cross-attention with queries of patches and keys of proxies. Such proxy $\rightarrow$ token absorption smooths the feature variation within a single part. Meanwhile, the part proxies absorb information from related patches using a semi-cross-attention and produce patch-absorbed proxies. Afterward, the proxy absorbs patch tokens within this part into the corresponding part descriptor (i.e., the right yellow box). The mutual absorbing strategy equips SAP with sufficient freedom to comply with the inherent inter-patch similarity out of the teacher supervision and contributes significantly to surpassing the teacher.

2) Supervision Signal Dropout. Since the teacher supervision is partially noisy and incorrect, it may confuse the partition learning as well as the part features. Hence, we further introduce dropping out part of them to produce sparse supervision in proxy-to-patch attention. Such signal dropout not only alleviates the target noise but also increases the learning freedom so that the student can learn to correct the teacher's errors. To eliminate redundant supervision, we tend to drop the supervision of recurring patches, that is, if a patch token is similar to others, it is likely to be abandoned. Specifically, we calculate a similarity matrix among all patches by the cosine distance between their token features, $\mathbf{T}_{srd} = \mathbf{I}^{'}\mathbf{I}^{'T} \in \mathbb{R}^{N\times N}$. And the dropout probability of each patch is positively correlated with its mean similarity to all other patches. Therefore, we randomly sample a certain number of $p_{drop}n$ patch tokens from the token sequence with a length of n, where p_{drop} is the dropout ratio. The probability weight of every patch is set to $p(i) = \text{softmax}(\sum \mathbf{T}_{srd}[i])$, where $[i]$ represents the i-th row of the matrix.

Experiments

Datasets and Evaluation Metrics

Occluded-DukeMTMC (Miao et al. 2019b) contains 15,618 training images, 17,661 gallery images, and 2,210 occluded query images, which is by far the largest occluded re-ID datasets. The experiments on this dataset follow the standard setting (Miao et al. 2019b) and the training, query, and gallery sets contain 9%, 100%, and 10% occluded images, respectively. **Occluded-REID** (Zhuo et al. 2018) consists of 2,000 images of 200 identities captured by mobile cameras. Each identity has five full-body gallery images and five occluded query images with different viewpoints and different types of severe occlusions. **Market-1501** (Zheng et al. 2015) consists of 32,668 images of 1,501 identities captured by 6 camera

views. Following the standard setting (Zheng et al. 2015), the whole dataset is divided into a training set containing 12,936 images of 751 identities and a testing set containing 19,732 images of 750 identities. **MSMT17** (Wei et al. 2018) contains 126,441 images of 4,101 IDs captured from a 15-camera network. The training set has 32,621 images of 1,041 identities, and the testing set has 93,820 images of 3,060 identities. During inference, 11,659 images are randomly selected as query and the other 82,161 images are used as gallery from the testing set (Wei et al. 2018).

Evaluation metric We adopt Cumulative Matching Characteristic (CMC) curve and mean average precision (mAP) for evaluations.

Implementation Details

We adopt ViT-Base pre-trained on ImageNet (Deng et al. 2009) as the global feature learning backbone, which contains 12 transformer layers. Based on this backbone, the student further appends a 2-layer 8-head transformer to conduct proxy-to-patch attention and to learn the part feature. In both training and inference stages, the person images are resized to 256×128 and the patch size is 16×16 with 4/5 pixels overlapping for holistic/occluded datasets (He et al. 2021). The training images are augmented with random horizontal flipping, random cropping and random erasing (Zhong et al. 2020) with a probability of 0.5. The batch size is set to 64 and the weight of distillation loss α is set to 1.0 for all datasets. The dropout ratio p_{drop} is set to 1/10 for MSMT17 and 1/8 to other three datasets. SGD optimizer is adopted with a momentum of 0.9 and the weight decay of 10^{-4}. The learning rate is initialized as 8×10^{-3} with cosine learning rate decay. The re-ID supervision consists of cross-entropy loss with label smoothing and softmax triplet loss. We use 1 NVIDIA A100 GPU for training.

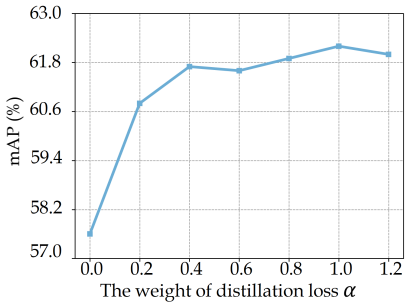

Figure 3: The improvements are more significant as α increases. The results are reported over Occluded-Duke dataset.

Comparison with the State-of-the-Art

We compare our method against the current state-of-the-art on four widely used occluded and holistic re-ID benchmarks in Table 1 and Table 2. Three categories of methods are compared, including methods based on inherent part-alignment, methods based on external cues and methods based on transformer. The external cue-based methods perform better than part-based ones in occluded scenarios due to the utilization of prior information of partition, while it is still limited by the noisy partition due to their inaccurate partition in ReID data. Our method outperforms them by large margin especially in occluded scenarios as in Table 1 because of the reasonable conflict that effectively resists the noisy external supervision. On the other hand, the transformer-based methods achieve excellent performance with self-attention, whereas it is still hard to learn discriminative partition automatically. With effective cooperation with the external teacher, SAP surpasses them by sufficiently using the prior partition information and resisting against the degradation by its noise. Our method also achieves superior performance in holistic re-ID demonstrating its generalization.

Method	O-Duke		O-REID	
	mAP	R-1	mAP	R-1
PCB (Sun et al. 2018)	33.7	42.6	38.9	41.3
DSR (He et al. 2018a)	30.4	40.8	62.8	72.8
SFR (He et al. 2018b)	32.0	42.3	-	-
FPR (He et al. 2019)	-	-	68.0	78.3
ISP (Zhu et al. 2020)	52.3	62.8	-	-
Part Bilinear (Suh et al. 2018)	-	36.9	-	-
FD-GAN (Ge et al. 2018)	-	40.8	-	-
PGFA (Miao et al. 2019a)	37.3	51.4	-	-
PVPM (Gao et al. 2020)	37.7	47.0	59.5	66.8
HONet (Wang et al. 2020)	43.8	55.1	70.2	80.3
GASM (He and Liu 2020)	-	-	65.6	74.5
RFCnet (Hou et al. 2021)	54.5	63.9	-	-
LKWS (Yang et al. 2021)	46.3	62.2	71.0	81.0
MoS (Jia et al. 2021)	55.1	66.6	-	-
PGFL-KD (Zheng et al. 2021)	54.1	63.0	70.3	80.7
PAT (Li et al. 2021)	53.6	64.5	72.1	81.6
Pirt (Ma, Zhao, and Li 2021)	50.9	60.0	-	-
DRL-Net (Jia et al. 2022)	53.9	65.8	-	-
AAformer (Zhu et al. 2021)	58.2	67.0	-	-
TransReID (He et al. 2021)	59.2	66.4	-	-
PFD (Wang et al. 2022a)	61.8	69.5	**83.0**	81.5
FED (Wang et al. 2022b)	56.4	68.1	79.3	**86.3**
SAP	**62.2**	**70.0**	76.8	83.0

Table 1: Comparison with state-of-the-art on occluded re-ID datasets, where O-Duke represents Occluded-Duke and O-REID represents Occluded-REID.

Ablation Studies

The effectiveness of semi-attention in SAP. Detailed ablation studies on occluded-Duke dataset are performed to evaluate the effectiveness of semi-attention for the SAP in Table 3. Line 1 denotes the baseline of our work without teacher supervision. According to Line 2 and Line 3 in Table 3, both patch-to-patch semi-attention and proxy-to-patch semi-attention improve the performance by applying them to the baseline individually. The improvement demonstrates that the two types of attention benefit from the cooperation and reasonable conflict from teacher-student learning. Moreover, combining the two types of semi-attention, the Line 4 achieves superior performance in their individual experi-

Method	Market-1501		MSMT17	
	mAP	R-1	mAP	R-1
PCB (Sun et al. 2018)	81.6	93.8	40.4	68.2
MGN (Wang et al. 2018)	86.9	95.7	52.1	76.9
VPM (Sun et al. 2019a)	80.8	93.0	-	-
ISP (Zhu et al. 2020)	88.6	95.3	-	-
SPReID (Kalayeh et al. 2018)	81.3	92.5	-	-
AANet (Tay, Roy, and Yap 2019)	82.5	93.9	-	-
P^2-Net (Guo et al. 2019)	85.6	95.2	-	-
PGFA (Miao et al. 2019a)	76.8	91.2	-	-
HONet (Wang et al. 2020)	84.9	94.2	-	-
GASM (He and Liu 2020)	84.7	95.3	52.5	79.5
RFCnet (Hou et al. 2021)	89.2	95.2	60.2	82.0
MoS (Jia et al. 2021)	89.0	95.4	-	-
PGFL-KD (Zheng et al. 2021)	87.2	95.3	-	-
PAT (Li et al. 2021)	88.0	95.4	-	-
Pirt (Ma, Zhao, and Li 2021)	86.3	94.1	-	-
DRL-Net (Jia et al. 2022)	86.9	94.7	55.3	78.4
AAformer (Zhu et al. 2021)	87.7	95.4	62.6	83.1
TransReID (He et al. 2021)	88.9	95.2	67.4	85.3
PFD (Wang et al. 2022a)	89.7	95.5	64.4	83.8
FED (Wang et al. 2022b)	86.3	95.0	-	-
SAP	**90.5**	**96.0**	**67.8**	**85.7**

Table 2: Comparison with state-of-the-art on holistic re-ID datasets Market-1501 and MSMT17. SAP is compatible to holistic re-ID and achieves competitive accuracy.

ments, indicating that the two attentions learn complementary information from the teacher.

The effectiveness of the conflict in SAP. We also evaluate the effectiveness of promoting reasonable conflict with the noisy teacher by removing each component, denoted as w/o. With the mutual absorbing partition refinement, Line 4 surpasses Line 5, indicating that the student learns additional inherent relation between the part-proxies and patch features by their mutually absorbing information from each other, and the refinement effectively gains resistance against the noisy supervision of the teacher. Similar results can be consistently observed in the comparison of Line 4 and Line 6, the supervision signal dropout can help learn more robust patch features by dropping the redundant and noisy supervision signals.

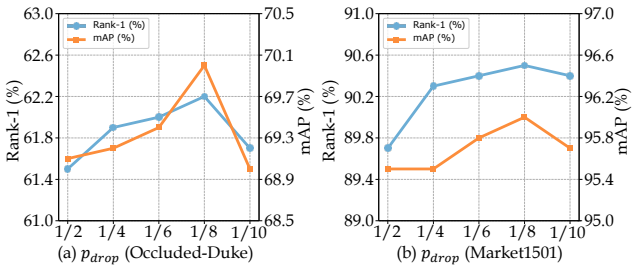

Figure 4: Evaluation of the performance with different dropout ratio p_{drop}.

Parameters analysis. Firstly, we study how the weight of distillation loss α in Eq. 5 and 7 affects learning student partition. Fig. 3 shows that the performance is improved as α increases as long as it is not too large, indicating the prior information of the noisy teacher is beneficial to the student for attention learning. Secondly, we study the impact of supervision signal dropout by using different dropout ratio p_{drop}. As shown in Fig. 4, SAP performs not well with a too large p_{drop} since useful supervision information is likely to be omitted. The best performance is achieved with $p_{drop} = 1/8$ consistently in Occluded-Duke and Market1501.

Methods	mAP	R-1	R-5
Base.	56.0	64.8	79.8
Base. +S-patch	59.6	66.6	81.0
Base. +S-proxy	60.1	66.4	81.2
Base. +S-patch +S-proxy	**62.2**	**70.0**	**83.0**
Base. +S-patch +S-proxy $_{w/o\ MAPR}$	60.8	68.0	81.2
Base. +S-patch +S-proxy $_{w/o\ SSD}$	61.8	68.9	82.4

Table 3: Ablation studies on Occluded-Duke dataset. Base., S-patch/proxy, MAPR, SSD denote baseline, proposed semi-supervised patch-/proxy-to-patch attention, mutual absorbing partition refinement and supervision signal dropout respectively.

Dataset	Occluded-Duke		Market-1501	
	mAP	R-1	mAP	R-1
ATR (Liang et al. 2015)	61.4	68.8	89.8	95.5
lip (Liang et al. 2018)	61.6	69.1	90.5	95.7
Pascal (Chen et al. 2014)	62.2	70.0	90.5	96.0

Table 4: Comparison of different datasets for human parsing teacher pre-training.

Different source datasets for teacher. We also evaluate our method with the best semantic partition teacher(human parsing) trained by different source data, including ATR, lip and pascal. All three datasets have domain gaps to the reID dataset making the trained teacher noisy. As shown in Table 4, SAP achieves the best results in pascal due to the smaller discrepancy of data. However, our method is effective in general even on the worst source data, demonstrating its practical in the real application.

Conclusion

This paper proposes a novel teacher-student strategy named semi-attention partition (SAP) for occluded re-ID. In SAP, the attention-based re-ID student learn to partition the person image from a noisy teacher and gains clear benefit. To this end, we integrate two forms of teacher-student consistency to reinforce their cooperation and a battery of well-engineered designs to promote reasonable conflict. Extensive experiments on re-ID benchmarks show that student brings general improvement after this semi-attention learning scheme and performs favorably against state-of-the-art methods.

References

Anil, R.; Pereyra, G.; Passos, A.; Ormandi, R.; Dahl, G. E.; and Hinton, G. E. 2018. Large scale distributed neural network training through online distillation. *arXiv preprint arXiv:1804.03235*.

Bagherinezhad, H.; Horton, M.; Rastegari, M.; and Farhadi, A. 2018. Label refinery: Improving imagenet classification through label progression. *arXiv preprint arXiv:1805.02641*.

Chen, X.; Mottaghi, R.; Liu, X.; Fidler, S.; Urtasun, R.; and Yuille, A. 2014. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1971–1978.

Cheng, X.; Jia, M.; Wang, Q.; and Zhang, J. 2022. More is better: Multi-source Dynamic Parsing Attention for Occluded Person Re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, 6840–6849.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Gao, S.; Wang, J.; Lu, H.; and Liu, Z. 2020. Pose-guided visible part matching for occluded person ReID. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 11744–11752.

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 (NIPS)*, 1230–1241.

Guo, J.; Yuan, Y.; Huang, L.; Zhang, C.; Yao, J.-G.; and Han, K. 2019. Beyond human parts: Dual part-aligned representations for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 3642–3651.

He, L.; Liang, J.; Li, H.; and Sun, Z. 2018a. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7073–7082.

He, L.; and Liu, W. 2020. Guided saliency feature learning for person re-identification in crowded scenes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 357–373.

He, L.; Sun, Z.; Zhu, Y.; and Wang, Y. 2018b. Recognizing Partial Biometric Patterns. *arXiv preprint arXiv:1810.07399*.

He, L.; Wang, Y.; Liu, W.; Zhao, H.; Sun, Z.; and Feng, J. 2019. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 8450–8459.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 15013–15022.

Hinton, G.; Vinyals, O.; Dean, J.; et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Hou, R.; Ma, B.; Chang, H.; Gu, X.; Shan, S.; and Chen, X. 2021. Feature completion for occluded person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jia, M.; Cheng, X.; Lu, S.; and Zhang, J. 2022. Learning Disentangled Representation Implicitly via Transformer for Occluded Person Re-Identification. *IEEE Transactions on Multimedia (TMM)*.

Jia, M.; Cheng, X.; Zhai, Y.; Lu, S.; Ma, S.; Tian, Y.; and Zhang, J. 2021. Matching on sets: Conquer occluded person re-identification without alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1673–1681.

Kalayeh, M. M.; Basaran, E.; Gökmen, M.; Kamasak, M. E.; and Shah, M. 2018. Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1062–1071.

Li, Y.; He, J.; Zhang, T.; Liu, X.; Zhang, Y.; and Wu, F. 2021. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2898–2907.

Liang, X.; Gong, K.; Shen, X.; and Lin, L. 2018. Look into person: Joint body parsing & pose estimation network and a new benchmark. 871–885.

Liang, X.; Liu, S.; Shen, X.; Yang, J.; Liu, L.; Dong, J.; Lin, L.; and Yan, S. 2015. Deep human parsing with active template regression. 2402–2414.

Lopez-Paz, D.; Bottou, L.; Schölkopf, B.; and Vapnik, V. 2015. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*.

Ma, Z.; Zhao, Y.; and Li, J. 2021. Pose-guided Inter-and Intra-part Relational Transformer for Occluded Person Re-Identification. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 1487–1496.

Miao, J.; Wu, Y.; Liu, P.; Ding, Y.; and Yang, Y. 2019a. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 542–551.

Miao, J.; Wu, Y.; Liu, P.; Ding, Y.; and Yang, Y. 2019b. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 542–551.

Suh, Y.; Wang, J.; Tang, S.; Mei, T.; and Lee, K. M. 2018. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European conference on computer vision (ECCV)*, 402–419.

Sun, Y.; Xu, Q.; Li, Y.; Zhang, C.; Li, Y.; Wang, S.; and Sun, J. 2019a. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 393–402.

Sun, Y.; Zheng, L.; Li, Y.; Yang, Y.; Tian, Q.; and Wang, S. 2019b. Learning part-based convolutional features for person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 43(3): 902–917.

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 *Proceedings of the 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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7134–7143.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 10347–10357. PMLR.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, G.; Yang, S.; Liu, H.; Wang, Z.; Yang, Y.; Wang, S.; Yu, G.; Zhou, E.; and Sun, J. 2020. High-order information matters: Learning relation and topology for occluded person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6449–6458.

Wang, G.; Yuan, Y.; Chen, X.; Li, J.; and Zhou, X. 2018. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, 274–282.

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

Wang, Z.; Zhu, F.; Tang, S.; Zhao, R.; He, L.; and Song, J. 2022b. Feature Erasing and Diffusion Network for Occluded Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4754–4763.

Wei, L.; Zhang, S.; Gao, W.; and Tian, Q. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 79–88.

Yang, C.; Xie, L.; Qiao, S.; and Yuille, A. L. 2019. Training deep neural networks in generations: A more tolerant teacher educates better students. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 5628–5635.

Yang, J.; Zhang, J.; Yu, F.; Jiang, X.; Zhang, M.; Sun, X.; Chen, Y.-C.; and Zheng, W.-S. 2021. Learning To Know Where To See: A Visibility-Aware Approach for Occluded Person Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11885–11894.

Yim, J.; Joo, D.; Bae, J.; and Kim, J. 2017. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4133–4141.

Yun, S.; Park, J.; Lee, K.; and Shin, J. 2020. Regularizing class-wise predictions via self-knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13876–13885.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9021–9030.

Zhai, Y.; Peng, P.; Jia, M.; Li, S.; Chen, W.; Gao, X.; and Tian, Y. 2022. Population-Based Evolutionary Gaming for Unsupervised Person Re-identification. *International Journal of Computer Vision*, 1–25.

Zhai, Y.; Ye, Q.; Lu, S.; Jia, M.; Ji, R.; and Tian, Y. 2020b. Multiple expert brainstorming for domain adaptive person re-identification. In *European Conference on Computer Vision*, 594–611. Springer.

Zhang, Y.; Xiang, T.; Hospedales, T. M.; and Lu, H. 2018. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4320–4328.

Zheng, K.; Lan, C.; Zeng, W.; Liu, J.; Zhang, Z.; and Zha, Z.-J. 2021. Pose-Guided Feature Learning with Knowledge Distillation for Occluded Person Re-Identification. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 4537–4545.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE 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*.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 13001–13008.

Zhu, K.; Guo, H.; Liu, Z.; Tang, M.; and Wang, J. 2020. Identity-Guided Human Semantic Parsing for Person Re-Identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 346–363.

Zhu, K.; Guo, H.; Zhang, S.; Wang, Y.; Huang, G.; Qiao, H.; Liu, J.; Wang, J.; and Tang, M. 2021. Aaformer: Auto-aligned transformer for person re-identification. *arXiv preprint arXiv:2104.00921*.

Zhuo, J.; Chen, Z.; Lai, J.; and Wang, G. 2018. Occluded person re-identification. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 1–6.