Diverse Knowledge Distillation for End-to-End Person Search

Xinyu Zhang[1,2], Xinlong Wang[2], Jia-Wang Bian[2], Chunhua Shen[2,3], Mingyu You[1]

[1]Tongji University, China
[2]The University of Adelaide, Australia
[3]Monash University, Australia
{zhangxinyu, myyou}@tongji.edu.cn, {xinlong.wang, jiawang.bian, chunhua.shen}@adelaide.edu.au

Abstract

Person search aims to localize and identify a specific person from a gallery of images. Recent methods can be categorized into two groups, *i.e.*, two-step and end-to-end approaches. The former views person search as two independent tasks and achieves dominant results using separately trained person detection and re-identification (Re-ID) models. The latter performs person search in an end-to-end fashion. Although the end-to-end approaches yield higher inference efficiency, they largely lag behind those two-step counterparts in terms of accuracy. In this paper, we argue that the gap between the two kinds of methods is mainly caused by the Re-ID subnetworks of end-to-end methods. To this end, we propose a simple yet strong end-to-end network with diverse knowledge distillation to break the bottleneck. We also design a spatial-invariant augmentation to assist model to be invariant to inaccurate detection results. Experimental results on the CUHK-SYSU and PRW datasets demonstrate the superiority of our method against existing approaches—it achieves on par accuracy with state-of-the-art two-step methods while maintaining high efficiency due to the single joint model. Code is available at: https://git.io/DKD-PersonSearch
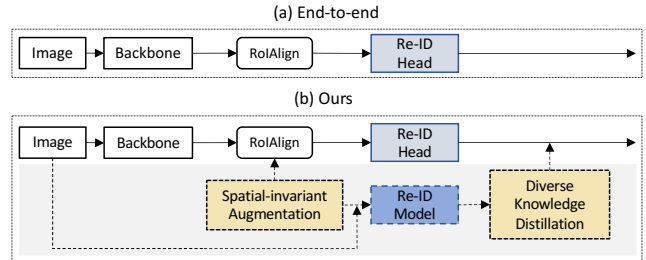
Figure 1: Structure comparison between (a) the plain end-to-end model and (b) our method for person search. Dashed lines are only involved during training. Our method maximizes the usage of guidance from the external Re-ID model via the proposed terms—see dashed boxes.

1 Introduction

Person search (Xu et al. 2014) is a recent hot topic in computer vision. The goal is to find a target person in a gallery of images, and hence involves both person detection and re-identification (Re-ID). Compared with solo Re-ID problem, person search faces greater challenges posed by detection, *e.g.*, inaccurate and misaligned bounding boxes may lead to poor identification.

Existing approaches for person search can be classified into two categories, *i.e.*, two-step strategy (Han et al. 2019; Lan, Zhu, and Gong 2018; Han et al. 2019; Wang et al. 2020a; Dong et al. 2020b) and end-to-end strategy (Xiao et al. 2017, 2019; Liu et al. 2017; Chang et al. 2018; He and Zhang 2018; Yan et al. 2019; Munjal, Amin, and Tombari 2019; Munjal, Galasso, and Amin 2019; Zhong, Wang, and Zhang 2020; Dong et al. 2020a; Chen et al. 2020a). Two-step approaches tackle the detection and Re-ID problem independently, in which the detection model crops the candidate persons from raw images, and the Re-ID model extracts fea-

tures for matching. These methods have shown high matching accuracy, but lag behind those end-to-end methods in efficiency due to the cascaded processing.

Thanks to the shared backbone for detection and Re-ID, the end-to-end methods boost the speed significantly. However, the performance is limited due to the conflict objectives of these two components in the joint learning framework (Chen et al. 2020a). Specifically, the detection head tends to find a common embedding space for all pedestrians, while the Re-ID head distinguishes different pedestrians with unique representations. Besides, feature maps of the proposals involve redundant context information (Zhu et al. 2019; Dong et al. 2020a), as they are pooled from the middle layer in the deep network, in which one pixel perceives a large area in the raw image. These issues consequently result in the inferior Re-ID accuracy.

In this paper, we boost the performance of end-to-end models towards efficient yet accurate person search. We first analyze the gap of both detection and Re-ID accuracy separately between two-step and end-to-end solutions. As shown in Table 1, the detection results are nearly the same in both settings, while there exists a large gap between final results (*i.e.*, Re-ID accuracy). Therefore, we argue that the Re-ID head is the main bottleneck of end-to-end models. We conjecture the reasons are twofold: (1) Two subtasks have conflicts when learning the shared features, which causes the learned features less discriminative for distinguishing fine-
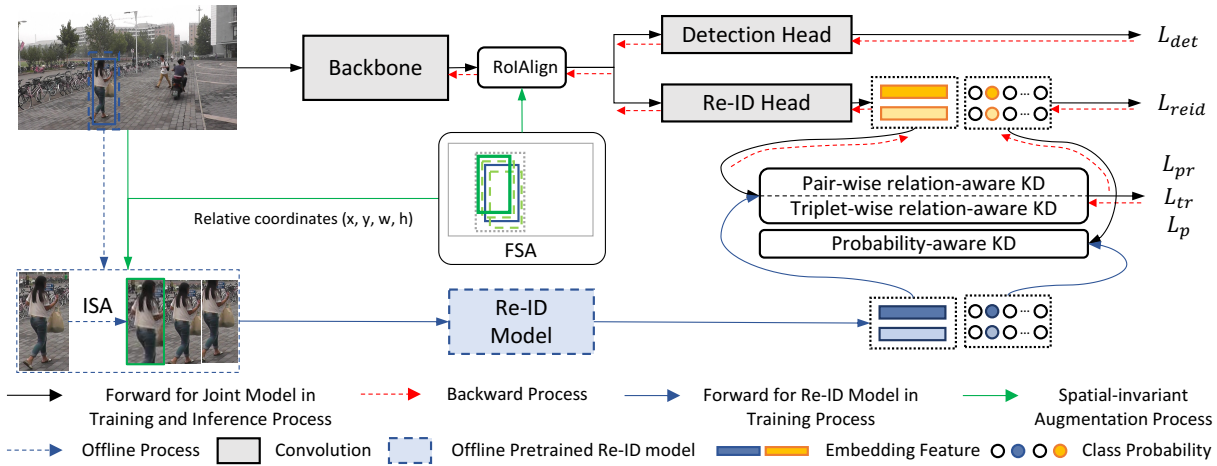
Figure 2: Overview of our person search framework. We first pre-train the external Re-ID model using our proposed image-level spatial-invariant augmentation (ISA). Then given the ground truth proposals, we use our proposed feature-level spatial-invariant augmentation (FSA) to obtain shifted bounding boxes in both images and feature maps. The internal Re-ID head and external Re-ID model take the shifted boxes in feature maps and images as input, respectively. Finally, we use three kinds of knowledge distillation (KD) to let the internal Re-ID head mimic the capability of the external Re-ID model, including the probability-aware KD and the proposed pair-wise as well as triplet-wise relation-aware KD.

grained persons; (2) The Re-ID head highly overfits the bounding box predictions, and inaccurate detection bounding box makes the matching fail, even it is a correct one (*i.e.*, IoU with ground truth is above 0.5).

To capitalize on our findings, we propose a simple yet effective person search framework for boosting the performance of the Re-ID head within the joint model[1]. The improvements come from two aspects. The first one is the enhancement on the output side with diverse knowledge distillation scheme. Since the Re-ID model in two-step approaches is trained on the person patches cropped from the raw images with ground truth bounding boxes, it can capture the discriminate features unaffected by the detection task. Inspired by knowledge distillation (KD) (Hinton, Vinyals, and Dean 2015), we make use of a separately trained Re-ID model to provide the superior prior knowledge to the Re-ID head. Apart from the traditional KD on the probability distribution, we design two relation-aware mimicking loss terms to make the Re-ID head fully emulate the sample relationships revealed by the well trained Re-ID model. The second one is a spatial-invariant augmentation mechanism on the input of the Re-ID head and the guiding Re-ID model. Instead of directly feeding ground truth person patches to the Re-ID head/model, we manually shift the cropping area of the proposals in the RoIAlign layer to mimic inaccurate detection results. To make the Re-ID head insensitive to this spatial disturbance, we present another KD guidance on the shifted proposals to minimize the difference between the probabilities produced by the Re-ID head and the Re-ID model. In this way, the Re-ID head is encouraged to improve the robustness on the detection variances.

Although external guidance is introduced, our method remains a lightweight framework since only the end-to-end

[1]Joint model is the alias of the end-to-end model in our paper.

model is reserved in the inference stage. Our contributions can be summarized as:

- We experimentally find that the main bottleneck of the end-to-end person search is the inferior Re-ID head.
- We propose a novel person search framework to enhance the Re-ID head via the diverse knowledge distillation and the spatial-invariant augmentation.
- We demonstrate the efficiency and efficacy of our method with comprehensive experiments, and show state-of-the-art results on the CUHK-SYSU and PRW datasets.

1.1 Related Work

Person search. Person search faces great challenges than person Re-ID (Zheng et al. 2015; Zheng, Zheng, and Yang 2017; Zhang et al. 2020b) due to unavailable perfect detection boxes. Existing methods in person search can be divided into two types, *i.e.*, two-step and end-to-end methods, according to performing the detection and Re-ID separately or jointly. Two-step methods train detectors (Tian et al. 2020; Ren et al. 2015) and Re-ID models separately and then combine them together in sequence (Zheng et al. 2017; Lan, Zhu, and Gong 2018; Chen et al. 2018; Han et al. 2019; Chen et al. 2018). (Zheng et al. 2017) evaluate various combinations of detectors and identifiers and propose a confidence weighted similarity method for person matching. (Lan, Zhu, and Gong 2018) study the multi-scale matching problem and propose a cross-level semantic alignment to address the problem. (Chen et al. 2018) enrich the feature representations by a mask-guided two-stream model. (Han et al. 2019) develop an ROI transform layer to refine sub-optimal boxes, in which the gradient flow is from the identifier. Recent work (Wang et al. 2020a) consider the consistency requirements between detection and Re-ID, and propose an

identity-guided query detector and a detection adapted re-ID model. These methods show that the two-step strategy can effectively alleviate the dilemma of detection and Re-ID and achieve high performance. Due to the separate processing of the two subtasks, these methods can be slow.

End-to-end methods are proposed to improve the efficiency by training a joint model. Among these methods, (Chang et al. 2018; Liu et al. 2017) directly search the person from the uncropped images by decision making or reinforcement learning. (Zhong, Wang, and Zhang 2020) address the misalignment issue and conduct robust partial matching. (Chang et al. 2018; Yan et al. 2019; Dong et al. 2020b,a; Chang et al. 2018) improve the feature discrimination with the guided query or surrounding persons. (Dong et al. 2020a) propose a bi-directional interaction network and utilize the cropped person patches as the guidance to reduce redundant context influences. Considering the inconsistency of detection and Re-ID, (Chen et al. 2020a) decompose the features of two subtasks in the polar coordinate system and propose a re-weighting method to focus on informative person areas.

In this work, we adopt the end-to-end strategy to remain high efficiency. We use an external Re-ID model to provide strong guidance via multiple knowledge distillation and propose the spatial-invariant augmentation method to improve the robustness of Re-ID head against inaccurate detections.

Guidance from external networks. It is a common practice to improve the main network using the additional guidance from external networks (Li, Jin, and Yan 2017; Cheng et al. 2018; Zhu et al. 2019; Hinton, Vinyals, and Dean 2015; Tung and Mori 2019; Liu et al. 2020b,a; Yuan et al. 2020; He et al. 2019). Two recent works in person search (Munjal, Galasso, and Amin 2019; Dong et al. 2020a) are related to our approach. (Munjal, Galasso, and Amin 2019) propose to distill the feature knowledge from a cumbersome model to a compact network on both detection and Re-ID subtasks, in which these two models solve the same person search task. (Dong et al. 2020a) use the guidance from the cropped person patches to eliminate the context influence outside the bounding boxes. Our proposed method aims to improve the discrimination of the Re-ID head by an independently trained Re-ID model. The optimization is supervised from diverse knowledge distillation mechanisms. Besides, our method makes the Re-ID head insensitive to inaccurate detections via a spatial-invariant augmentation scheme.

2 Method

In this section, we first conduct a comparative study to show that the Re-ID head is the main bottleneck in the end-to-end model (Section 2.1). We then present the proposed person search framework, as illustrated in Figure 2. The overall structure can provide powerful guidances tailored for both input and output of the networks. For output, we propose the diverse knowledge distillation scheme (Section 2.2) to extract multiple guidances accessed from the external Re-ID model, in which consisting of the probability-aware knowledge distillation (KD) and the relation-aware knowledge distillation. As for input, a spatial-invariant augmentation method (Section 2.3) is introduced to make the Re-ID head

| Strategy | CUHK-SYSU | | | | PRW | | | |
| | Detection | | Re-ID | | Detection | | Re-ID | |
	AP	Recall	mAP	top-1	AP	Recall	mAP	top-1
two-step	87.02	89.69	91.97	93.14	92.39	94.08	51.85	85.37
end-to-end	87.23	90.02	86.47	87.55	90.85	92.93	41.91	82.11
Ours	87.87	90.62	93.09	94.24	90.71	92.71	50.51	87.07

Table 1: Component performance comparison between the end-to-end and two-step methods. Model structures are described in Section 3.3.

robust to the noisy detection results.

2.1 Gap Between Two-Step and End-to-End Strategies

Two first rows in Table 1 are the results of end-to-end and two-step methods[2]. It shows that the detection results in both settings are similar, especially in the CUHK-SYSU dataset, *i.e.*, 87.02% *vs.* 87.23% in the AP metric. However, there is a large gap in their final matching results. Specifically, the two-step strategy outperforms the end-to-end strategy by 5.5% and 9.9% mAP on CUHK-SYSU and PRW datasets, respectively. This motivates us to regard the Re-ID head as the main bottleneck of end-to-end person search models. We propose our novel methods in the following sections to improve the capability of the Re-ID head and bridge the gap between the two kinds of strategies.

2.2 Diverse Knowledge Distillation

As indicated in Table 1, the separate Re-ID model in the two-step strategy is able to accurately distinguish fine-grained pedestrians, which inspires us to introduce an external well-trained Re-ID model (denoted as $\mathcal{M}$) to reinforce our Re-ID head (denoted as $\mathcal{H}$) in the end-to-end model. The direct way is to apply knowledge distillation (KD) to transfer the informative knowledge from $\mathcal{M}$ to $\mathcal{H}$. In addition to the traditional KD, *i.e.*, probability-aware KD, we introduce a relation-aware KD scheme which consists of the pair-wise and the triplet-wise knowledge distillation.

Probability-aware knowledge distillation. Following the conventional KD (Hinton, Vinyals, and Dean 2015), we employ a straightforward guidance to let head $\mathcal{H}$ mimic $\mathcal{M}$ by minimizing the Kullback-Leibler (KL) divergence between the predicted class probabilities of the two networks. The output probability from $\mathcal{M}$ can be regarded as the soft target to serve as the informative guidance. The loss function can be written as:

$$L_p = \frac{1}{B} \sum_{i=1}^{B} \mathrm{KL}(\mathbf{p}_i^{\mathcal{H}} \parallel \mathbf{p}_i^{\mathcal{M}}), \qquad (1)$$

---

[2]For the two-step method, we train the detection and the Re-ID model individually. Bounding boxes are first produced by the detector and then passed through the Re-ID model to obtain the feature embeddings in the inference. For the end-to-end method, the joint model with the detection and Re-ID head is simultaneously trained and the feature embeddings are extracted via inputting the images to the joint model directly (details in Section 3.2).
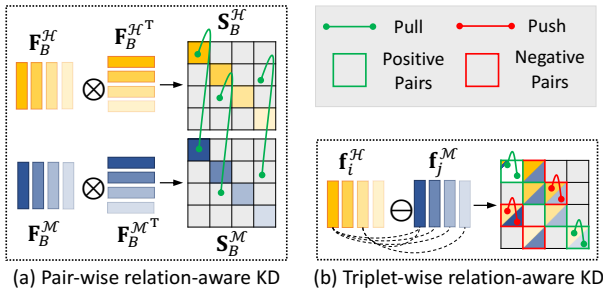
(a) Pair-wise relation-aware KD   (b) Triplet-wise relation-aware KD

Figure 3: Illustration of our proposed (a) pair-wise and (b) triplet-wise relation-aware KD. $\mathcal{H}$ represents the Re-ID head in the end-to-end model, while $\mathcal{M}$ is the external Re-ID model. Both two relation guidances push the Re-ID head to mimic the sample relationships of the Re-ID model.

where $\mathbf{p}_i^{\mathcal{H}}$ and $\mathbf{p}_i^{\mathcal{M}}$ denotes the probability output from $\mathcal{H}$ and $\mathcal{M}$ of the i-th sample in a mini-batch B. $\mathrm{KL}(\cdot)$ denotes KL divergence between two probabilities.

Relation-aware knowledge distillation. Some works have already verified the success of discovering sample relationships in person Re-ID task (Luo et al. 2019; He et al. 2020b; Hermans, Beyer, and Leibe 2017; Zhang et al. 2020a; Wang et al. 2020b; Zhang et al. 2019). It motivates us to transfer the informative relation knowledge among samples from $\mathcal{M}$ to $\mathcal{H}$. Inspired by (Tung and Mori 2019; Liu et al. 2020b), we present the *pair-wise relation-aware knowledge distillation* to mimic the similarity distribution between sample pairs. As shown in Figure 3 (a), for a specific mini-batch, let $\mathbf{S}_B^{\mathcal{H}}$ and $\mathbf{S}_B^{\mathcal{M}}$ represent the similarity matrix among all sample pairs from the head $\mathcal{H}$ and model $\mathcal{M}$, respectively. In detail, $\mathbf{S}_B^{\mathcal{H}}$ is produced by a row-wise L2 normalization on the outer product of the mini-batch embedding feature $\mathbf{F}_B^{\mathcal{H}}$:

$$\mathbf{S}_B^{\mathcal{H}} \leftarrow \mathbf{F}_B^{\mathcal{H}} \otimes \mathbf{F}_B^{\mathcal{H}^{\mathrm{T}}}; \mathbf{S}_{B\{i\}}^{\mathcal{H}} \leftarrow \mathbf{S}_{B\{i\}}^{\mathcal{H}}/||\mathbf{S}_{B\{i\}}^{\mathcal{H}}||_2, \quad (2)$$

where $\{i\}$ represents the i-th row in the similarity matrix. The same process is applied on $\mathbf{S}_B^{\mathcal{M}}$ from the embedding feature $\mathbf{F}_B^{\mathcal{M}}$ in the model $\mathcal{M}$. Then we apply the element-wise mean squared loss function on $\mathbf{S}_B^{\mathcal{H}}$ and $\mathbf{S}_B^{\mathcal{M}}$:

$$L_{pr} = \frac{1}{B^2}||\mathbf{S}_B^{\mathcal{H}} - \mathbf{S}_B^{\mathcal{M}}||_2^2, \quad (3)$$

In this way, the head $\mathcal{H}$ can capture the similar sample relationships as the model $\mathcal{M}$.

Motivated by recent contrastive learning algorithms (Chen et al. 2020b; He et al. 2020a; Tian, Krishnan, and Isola 2020), there exists a self supervision between differently generated views of the same data. In our method, for a given data x_i, the feature embedding from $\mathcal{H}$ and $\mathcal{M}$, denoted as $\mathbf{f}_i^{\mathcal{H}}$ and $\mathbf{f}_i^{\mathcal{M}}$, can be viewed as a positive pair. Instead, we treat any other samples within a mini-batch as negative examples. The corresponding feature embedding $\mathbf{f}_i^{\mathcal{H}}$ and $\mathbf{f}_j^{\mathcal{M}}$ is thus a negative pair, where $i, j \in B, i \neq j$. Following the above assumption, we propose the *triplet-wise relation-aware knowledge distillation* to maximize the agreement of different feature representations of the same

individual from $\mathcal{H}$ and $\mathcal{M}$, while minimizing that from different persons. The loss function is defined as:

$$L_{tr} = \frac{1}{B} \sum_{i=1}^{B} [m + ||\mathbf{f}_i^{\mathcal{H}} - \mathbf{f}_i^{\mathcal{M}}||_2 - \min_{\substack{i \neq j \\ i,j \in B}} ||\mathbf{f}_i^{\mathcal{H}} - \mathbf{f}_j^{\mathcal{M}}||_2]_+,$$

$$(4)$$

where m is the margin between positive and negative pairs. Here, we mine the hardest negative sample $\mathbf{f}_j^{\mathcal{M}}$ within a mini-batch B as in (Hermans, Beyer, and Leibe 2017).

We optimize the joint model with both of the probability-aware and the relation-aware KD to better capture sample probabilities and correlations from the external well-trained Re-ID model. The whole objective function is presented as:

$$L = L_{det} + \lambda(L_{reid} + \beta_p L_p + \beta_{pr} L_{pr} + \beta_{tr} L_{tr}), \quad (5)$$

where L_{det} and L_{reid} are the original detection and Re-ID loss functions. Here we use softmax cross-entropy loss for L_{reid}. λ acts as the weight to balance the detection and Re-ID losses. β_p, β_{pr} and β_{tr} are the loss weights for different types of guidances from $\mathcal{M}$, respectively.

2.3 Spatial-Invariant Augmentation

In person Re-ID community, a major research line is to deal with the misalignment problem caused by occlusions, partial bodies or wrongly detected boxes (Sun et al. 2019; Guo et al. 2019; Sun et al. 2018; Zhao et al. 2020; Miao et al. 2019). The goal of existing methods aims to learn a robust model to extract informative features from various misaligned persons. In the person search task, the quality of Re-ID features is sensitive to the detection results to some extent. To this end, we propose a spatial-invariant augmentation mechanism at both image level and feature level.

Image-level spatial-invariant augmentation. In (Chen et al. 2018), authors show that the expanded RoI works better for Re-ID model. However, too much background will reduce the performance. Instead of directly using the expanded images, we propose the image-level spatial-invariant augmentation (ISA) for model $\mathcal{M}$ to stably improve the capability of the external Re-ID guidance. In detail, we expand the ground truth bounding boxes by a pixel length α, resulting in loose boxes, in which they share the same centroids with the corresponding ground truths. We then use the loose boxes with the random crop operation to train the Re-ID model, denoted as $\widetilde{\mathcal{M}}$. Considering that there are multiple boxes with various scales, we need to impose an upper-bound restriction on the range of the random crop operation according to different box sizes. In Re-ID, we usually add zero padding with Δp pixels and then crop and resize the images to a fixed size $H \times W$. Based on that, for a specific ground truth box with height h and width w, the expanded range of them can be adapted to $r_h = \min(2h \times \Delta p/H, \alpha)$ and $r_w = \min(2w \times \Delta p/W, \alpha)$ respectively. $\min$ operation is to prevent the cropped boxes far from the ground truths. In this way, the external model boosts the spatial-invariant ability by simulating inaccurate boxes.

Feature-level spatial-invariant augmentation. In addition to ISA on the external $\widetilde{\mathcal{M}}$, we also improve the robustness of

the Re-ID head $\mathcal{H}$ against inaccurate detection boxes. Under the assumption that $\widetilde{\mathcal{M}}$ is good enough to recognize the identities even if the generated bounding boxes are not well aligned to ground truths, we propose a feature-level spatial-invariant augmentation (FSA) to let $\mathcal{H}$ mimic the discriminative capability of $\widetilde{\mathcal{M}}$ on spatial variances. Specifically, given a ground truth proposal, we perform the RoIAlign (He et al. 2017) in the range of the proportional pixels *w.r.t.* pixel range in ISA, *i.e.*, r_w/S and r_h/S. Here, S is the stride pixel of the current feature map *w.r.t.* the input image. Consequently, there is an additional probability $\tilde{\mathbf{p}}_i^{\mathcal{H}}$ with the spatial variance for a given data x_i.

For better capturing the consistent information from $\widetilde{\mathcal{M}}$, we also extract the probability $\tilde{\mathbf{p}}_i^{\mathcal{M}}$ from $\widetilde{\mathcal{M}}$ by cropping the person patch with the relative coordinate on the raw image *w.r.t.* the above spatial variance. Then the objective function of the probability-aware KD in Section 2.2 is re-defined as:

$$L_p = \frac{1}{B} \sum_{i=1}^{B} [\mathrm{KL}(\mathbf{p}_i^{\mathcal{H}} \parallel \mathbf{p}_i^{\widetilde{\mathcal{M}}}) + \mathrm{KL}(\tilde{\mathbf{p}}_i^{\mathcal{H}} \parallel \tilde{\mathbf{p}}_i^{\widetilde{\mathcal{M}}})], \quad (6)$$

where the original $\mathcal{M}$ is replaced by $\widetilde{\mathcal{M}}$ after ISA and L_p is updated from Eq. (5) to Eq. (6). By this means, $\mathcal{H}$ is relatively insensitive to various detection results.

3 Experiments

3.1 Evaluation Settings

Datasets. We carry out experiments on CUHK-SYSU and PRW datasets. CUHK-SYSU (Xiao et al. 2017) is a large-scale dataset extracted from the street and movie snapshots. It consists of 18,184 images with 96,143 annotated bounding boxes with 8,432 labeled identities. The training set contains 11,206 images with 5,532 identities, while the test set includes 6,978 gallery images and 2,900 query persons. A set of protocols with gallery size ranging from 50 to 4000 is defined to evaluate the model scalability. The gallery size is set to 100 by default without specificity. PRW dataset (Zheng et al. 2017) is collected by six cameras at different locations on a university campus. A total of 11,816 frames are annotated with 43,110 bounding boxes, among which 34,304 are assigned labels from 1 to 932, and the rest marked unknown identities. The training set consists of 5,704 frames with 483 identities, and the test set contains 6,112 gallery images and 2,057 query images with 450 identities.

Metrics. We use the standard evaluation metrics in person search (Xiao et al. 2017), *i.e.*, cumulative matching characteristics (CMC) and mean average precision (mAP). CMC is inherited from the person Re-ID problem showing the probabilities that at least one of the top-K predicted bounding boxes overlaps with the ground truths with IoU greater or equal to 0.5. mAP is inspired by object detection, where the average precision (AP) for each query is computed from the area under its precision-recall curve and then mAP is calculated by averaging APs across all queries. For detection, we utilize Recall and Average Precision (AP) as the detection metrics following (Chen et al. 2020a).

Strategy	Method	Augmentation Type		PRW	
		ISA	FSA	mAP	top-1
two-step	[D; R-1]	-	-	51.85	85.37
	[D; R-2]	✓	-	54.16	87.89
	[GT; R-1]	-	-	54.46	86.63
	[GT; R-2]	✓	-	56.78	89.30
end-to-end	Ours-0	-	-	41.91	82.11
	Ours-1	-	-	49.44	85.85
	Ours-2	✓	-	49.78	86.19
	Ours-3	✓	✓	50.15	86.53

Table 2: Efficacy of the proposed image-level augmentation (ISA) and the feature-level augmentation (FSA) on PRW dataset. [X;Y] represents the cascaded operation of X and Y model. Full notations are described in Section 3.3.

Strategy	Method	Type			PRW	
		L_{pr}	L_{tr}	L_p	mAP	top-1
two-step	[D; R-2]	-	-	-	54.16	87.89
end-to-end	Ours-0	-	-	-	41.91	82.11
	Ours-4	✓	-	-	46.00	84.59
	Ours-5	-	✓	-	45.54	83.37
	Ours-6	✓	✓	-	47.18	84.83
	Ours-7	✓	-	✓	50.18	86.78
	Ours-8	-	✓	✓	50.46	86.58
	Ours	✓	✓	✓	**50.51**	**87.07**

Table 3: Efficacy of different supervision guidance on PRW.

3.2 Implementation Details

Model. For the end-to-end model in our method, we adopt Faster R-CNN (Ren et al. 2015) as our backbone network, in which ResNet-50 (He et al. 2016) pretrained on ImageNet (Deng et al. 2009) is used. As in (Chen et al. 2020a; Dong et al. 2020a), the stem network has four residual blocks from conv1 to conv4. A standard RPN is built on the stem network to generate candidate proposals, followed by the RoIAlign (He et al. 2017) operation to crop and reshape the proposals to 16×8. The detection head and the Re-ID head are two separate conv5 residual blocks without parameter sharing. The task-specific heads help alleviate the conflicts across tasks with a small overhead in computation. Following (Luo et al. 2019), we adopt a batch-normalization layer before the last fully-connection layer. The dimension of feature embeddings from the Re-ID head is set to 2048. For the external Re-ID model, we employ a popular baseline (Luo et al. 2019) based on the ResNet-50 backbone.

Training. We train all models using SGD optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4}. Input images are resized to have at least 800 pixels on the short size and at most 1333 pixels on the long side. We set the batch size to 4. The initial learning rate is 0.001 and then multiplied by 0.1 after 3×10^4 iterations. The total iteration is 5×10^5. We set $\lambda = 0.1$, $\beta_p = 0.1$ $\beta_{pr} = \beta_{tr} = 1.0$ and $m = 0.3$ in Eq. (4) and S to 16 in FSA. For the Re-ID model, we use the softmax cross-entropy loss and the SGD optimizer. We resize all the person patches to $H \times W =$

	Method	CUHK-SYSU		PRW	
		mAP	top-1	mAP	top-1
Two-step	MGTS (Chen et al. 2018)	83.0	83.7	32.6	72.1
	CLSA (Lan, Zhu, and Gong 2018)	87.2	88.5	38.7	65.0
	RDLR (Han et al. 2019)	93.0	94.2	42.9	70.2
	TCTS (Wang et al. 2020a)	**93.9**	**95.1**	46.8	87.5
	IGPN+PCB (Dong et al. 2020b)	90.3	91.4	47.2	87.0
	[D; R-2] (Ours two-step)	93.60	94.72	**54.16**	**87.89**
End-to-end	OIM (Xiao et al. 2017)	75.5	78.7	21.3	49.9
	IAN (Xiao et al. 2019)	76.3	80.1	23.0	53.1
	NPSM (Liu et al. 2017)	77.9	81.2	24.2	53.1
	RCAA (Chang et al. 2018)	79.3	81.3	-	-
	I-Net (He and Zhang 2018)	79.5	81.5	-	-
	CTXGraph (Yan et al. 2019)	84.1	86.5	33.4	73.6
	QEEPS*	88.9	89.1	37.1	76.7
	QEEPS+*	85.0	85.5	39.7	80.0
	APNet*	88.9	89.3	41.9	81.4
	BINet (Dong et al. 2020a)	90.0	90.7	45.3	81.7
	NAE (Chen et al. 2020a)	91.5	92.4	43.3	80.9
	NAE+ (Chen et al. 2020a)	92.1	92.9	44.0	81.1
	Ours-0 (Ours baseline)	86.47	87.55	41.91	82.11
	Ours (Ours full)	**93.09**	**94.24**	**50.51**	**87.07**

Table 4: Person search results on CUHK-SYSU and PRW datasets. The gallery size of CUHK-SYSU is 100. * denotes that QEEPS, QEEPS+ and APNet are from (Munjal, Amin, and Tombari 2019), (Munjal, Galasso, and Amin 2019) and (Zhong, Wang, and Zhang 2020).

256×128. The batch size is set to 64. The initial learning rate is 3.5×10^{-4} and the total training epoch is 240. Besides, we set Δp to 10 and α to 40 in the process of ISA. In the inference stage, we only keep the joint model.

3.3 Ablation Study

For simplicity, we first introduce the following variants:

- D: the detection model only with L_{det} in Eq. (5).
- R-1: the Re-ID model $\mathcal{M}$ as described in Section 3.2. The model is trained with person patches with labeled ground truth bounding boxes.
- R-2: adding the proposed ISA on person patches compared with R-1.
- Ours-0: the plain model trained in the end-to-end strategy. Details are described in Section 3.2. L_{det} and L_{reid} are used for optimization in Eq. (5).
- Ours-1: adding the probability-aware knowledge distillation (KD) from R-1 on Ours-0. L_{det}, L_{reid} and L_p are loss functions in Eq. (5).
- Ours-2: adding the probability-aware KD from R-2 on Ours-0.
- Ours-3: adding the proposed FSA on Ours-2, *i.e.*, replacing L_p with Eq. (6) in Eq. (5).
- Ours-4 / -5: adding the proposed pair-wise / triplet-wise relation-aware KD from R-2 on Ours-0. L_{det}, L_{reid} and L_{pr} / L_{tr} are loss functions in Eq. (5).
- Ours-6: adding both pair-wise and triplet-wise relation-aware KD from R-2 on Ours-0 simultaneously. L_{det}, L_{reid}, L_{pr} and L_{tr} are loss functions in Eq. (5).
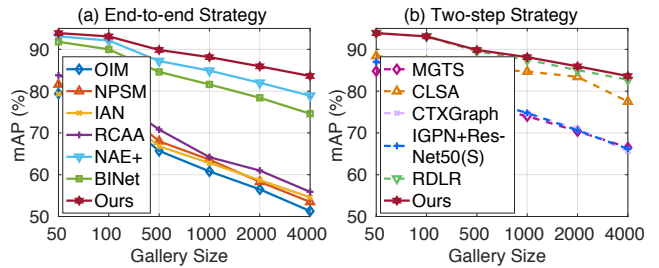

Figure 4: CUHK-SYSU results with varying gallery sizes.

- Ours-7 / -8: adding probability-aware KD in Eq. (6) on Ours-4 / -5. L_p is added as the loss function in Eq. (5).
- Ours: adding probability-aware KD in Eq. (6) on Ours-6, which is the whole setting of our method.

Effectiveness of ISA and FSA. Table 2 shows the ablation study results on PRW dataset. The results ([D; R-1] *vs.* [D; R-2] and Ours-1 *vs.* Ours-2) demonstrate the efficacy of our image-level spatial-invariant augmentation (ISA). It shows that ISA leads to 2.31% mAP improvement in the two-step setting. When applying the Re-ID model with ISA as the external guidance in our method, there are 0.34% mAP and 0.34% top-1 improvements. It is because the Re-ID model with ISA can provide a robust external guidance when detection results are inaccurate. Moreover, in Ours-3, we apply the feature-level spatial-invariant augmentation (FSA) on Ours-2, resulting in further improvement of the performance to 50.15% mAP and 86.53% top-1. It is because that the Re-ID head can effectively learn the spatial invariance with the help of FSA, so that the influences from the detection head can be suppressed to some extent.

Effectiveness of different knowledge distillation schemes. We evaluate the impact of different KD schemes as described in Section 2.2. Results are performed on PRW dataset and shown in Table 3, from which we can observe that: i) Our Re-ID head successfully learns the discriminative ability under the guidance from the external Re-ID model. With all three types of the supervision guidances, our method improves 8.6% mAP and 4.96% top-1, which is close to the results of two-step strategy. ii) Either pair-wise or triplet-wise relation-aware KD can improve the performance and the result is further improved with both of them. We believe that it is because our Re-ID head well captures reliable sample relationships provided by the Re-ID model which improves the feature discrimination.

3.4 Comparison with State-of-the-art Methods

Following the evaluation setting in (Xiao et al. 2017; Zheng et al. 2017; Chen et al. 2020a), we compare our proposed method with state-of-the-art methods on person search. As shown in Table 4, all methods are categorized into two parts, *i.e.*, two-step strategies in the upper block and the end-to-end methods in the lower block.

Results on CUHK-SYSU. As shown in Table 4, our method achieves the best results in the end-to-end category. Com-
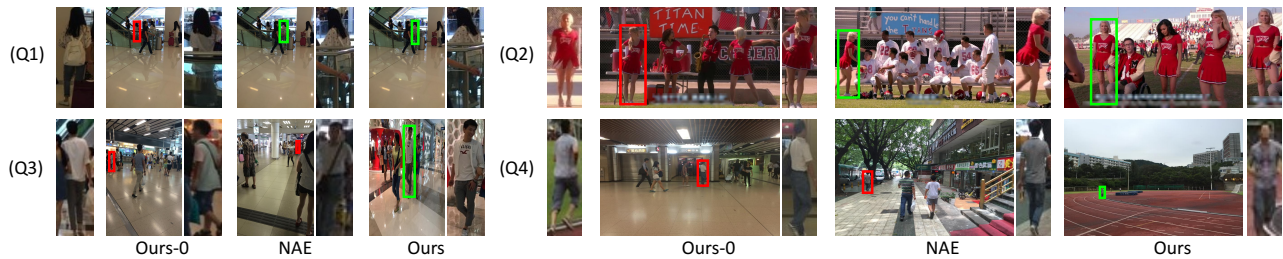
Figure 5: Sample visual results. For each query 'Q', we show the top-1 search result by our baseline (Ours-0), NAE (Chen et al. 2020a) and the our whole method (Ours). Green and red rectangles denote the correct and wrong results respectively.

Image Size	End-to-end			Two-step
	Ours	NAE	BINet	[D; R-2]
900x1500	124	141	80	175

Table 5: Running time (ms) comparison. [D; R-2] stands for the two-step model cascaded by detection and Re-ID.

Re-ID model	mAP	top-1
Ours	50.51	87.07
w/o random erasing (Zhong et al. 2020)	50.19	86.58
w/o BNNeck (Luo et al. 2019)	47.83	85.03

Table 6: Effect of different external Re-ID models on PRW.

pared with NAE+ (Chen et al. 2020a) that decomposes the detection and Re-ID subtasks via a norm-aware embedding, our method focuses on the improvement of the Re-ID head guided by the external Re-ID model and results in the superior performance (93.09% *vs.* 92.1% mAP). Our method also outperforms another prior best method BINet (Dong et al. 2020a) by 1.59% mAP. Instead of alleviating the negative context information as in (Dong et al. 2020a), our method aims to learn spatial-invariant features that are insensitive to inaccurate detection results. Moreover, our proposed relation-aware KD explores the sample relationships, which is important to the Re-ID task. From Figure 5, top-1 search results from the baseline, NAE (Chen et al. 2020a) and our method are visualized.

We also evaluate the scalability of different methods with various gallery sizes, *i.e.*, {50, 100, 500, 1000, 2000, 4000}. As shown in Figure 4, all methods degrade the performance monotonically as the gallery size increases. It shows that person search in large search scopes remains a challenging task as it involves more distractors. Nevertheless, our method outperforms all approaches, especially under large gallery sizes. There is only a relatively small drop in mAP when the gallery size changes from 50 to 4000, which demonstrates the scalability of our approach. Interestingly, our method achieves comparable results and even surpasses all two-step methods with the gallery size increasing, which shows that our method can improve the model discrimination and obtain more spatial-invariant features.

Results on PRW. As shown in Table 4, our baseline model is already competitive with other end-to-end methods. This

indicates that it would be better to separate the detection and Re-ID heads in the end-to-end model without parameter sharing. Overall, our method surpasses all the methods both in the end-to-end and two-step strategy except TCTS (Wang et al. 2020a) at top-1. Notably, it outperforms the best end-to-end method BINet (Dong et al. 2020a) by 5.21% mAP, showing the superiority of our method.

Running time. We report the speed of different approaches in Table 5. All experiments are conducted on a GeForce GTX 1080 Ti machine. For fairness, the size of input images is set to 900×1500. We reproduce the result of NAE (Chen et al. 2020a) on the same machine, while reporting the original result of BINet (Dong et al. 2020a) since its code is not available. Table 5 shows that our method is faster than NAE. Note that our method can achieve the same efficiency when applying the same backbone as BINet, as no additional parameter is introduced. Moreover, our method is more efficient than the two-step strategy (*i.e.*, [D; R-2]).

Discussion. In order to show the effectiveness of our method clearly, we evaluate the effect of different external Re-ID models. As shown in Table 6, when removing the important schemes in the Re-ID model, *i.e.*, random erasing (Zhong et al. 2020) and BNNeck (Luo et al. 2019), the performance is slightly lower than the full model yet outperforms all the recent methods. We also explore the detection results influenced by our method in Table 1. Compared with the baseline, our method has limited influence on the detection part.

4 Conclusion

In this paper, we identify that the Re-ID head is the main bottleneck in the end-to-end person search model. To this end, we propose a strong person search network to improve the Re-ID head with an external Re-ID model. We provide guidance on both the output and input of the model. For the model output, we design a diverse knowledge distillation, consisting of probability-aware and relation-aware KD, to let the Re-ID head mimic the well-trained Re-ID model. For the model input, we develop an image-level and feature-level spatial-invariant augmentations to make the Re-ID head insensitive to inaccurate detection results. Only the end-to-end model is needed during inference so that there is no additional computation. Extensive experiments show the effectiveness of our methods.

References

Chang, X.; Huang, P.-Y.; Shen, Y.-D.; Liang, X.; Yang, Y.; and Hauptmann, A. G. 2018. RCAA: Relational context-aware agents for person search. In *Proc. Eur. Conf. Comp. Vis.*, 84–100.

Chen, D.; Zhang, S.; Ouyang, W.; Yang, J.; and Tai, Y. 2018. Person search via a mask-guided two-stream cnn model. In *Proc. Eur. Conf. Comp. Vis.*, 734–750.

Chen, D.; Zhang, S.; Yang, J.; and Schiele, B. 2020a. Norm-Aware Embedding for Efficient Person Search. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 12615–12624.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020b. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709* .

Cheng, B.; Wei, Y.; Shi, H.; Feris, R.; Xiong, J.; and Huang, T. 2018. Revisiting rcnn: On awakening the classification power of faster rcnn. In *Proc. Eur. Conf. Comp. Vis.*, 453–468.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 248–255. Ieee.

Dong, W.; Zhang, Z.; Song, C.; and Tan, T. 2020a. Bi-Directional Interaction Network for Person Search. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2839–2848.

Dong, W.; Zhang, Z.; Song, C.; and Tan, T. 2020b. Instance Guided Proposal Network for Person Search. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2585–2594.

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 *Proc. IEEE Int. Conf. Comp. Vis.*, 3642–3651.

Han, C.; Ye, J.; Zhong, Y.; Tan, X.; Zhang, C.; Gao, C.; and Sang, N. 2019. Re-id driven localization refinement for person search. In *Proc. IEEE Int. Conf. Comp. Vis.*, 9814–9823.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020a. Momentum contrast for unsupervised visual representation learning. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 9729–9738.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 770–778.

He, L.; Liao, X.; Liu, W.; Liu, X.; Cheng, P.; and Mei, T. 2020b. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631* .

He, T.; Shen, C.; Tian, Z.; Gong, D.; Sun, C.; and Yan, Y. 2019. Knowledge adaptation for efficient semantic segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 578–587.

He, Z.; and Zhang, L. 2018. End-to-end detection and re-identification integrated net for person search. In *Proc. Asian Conf. Comp. Vis.*, 349–364. Springer.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* .

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* .

Lan, X.; Zhu, X.; and Gong, S. 2018. Person search by multi-scale matching. In *Proc. Eur. Conf. Comp. Vis.*, 536–552.

Li, Q.; Jin, S.; and Yan, J. 2017. Mimicking very efficient network for object detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 6356–6364.

Liu, H.; Feng, J.; Jie, Z.; Jayashree, K.; Zhao, B.; Qi, M.; Jiang, J.; and Yan, S. 2017. Neural person search machines. In *Proc. IEEE Int. Conf. Comp. Vis.*, 493–501.

Liu, Y.; Shen, C.; Yu, C.; and Wang, J. 2020a. Efficient Semantic Video Segmentation with Per-frame Inference. In *Proc. Eur. Conf. Comp. Vis.*, 352–368. Springer.

Liu, Y.; Shu, C.; Wang, J.; and Shen, C. 2020b. Structured knowledge distillation for dense prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* .

Luo, H.; Gu, Y.; Liao, X.; Lai, S.; and Jiang, W. 2019. Bag of tricks and a strong baseline for deep person re-identification. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn. Workshop*.

Miao, J.; Wu, Y.; Liu, P.; Ding, Y.; and Yang, Y. 2019. Pose-guided feature alignment for occluded person re-identification. In *Proc. IEEE Int. Conf. Comp. Vis.*, 542–551.

Munjal, B.; Amin, S.; and Tombari, F. 2019. Query-guided end-to-end person search. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 811–820.

Munjal, B.; Galasso, F.; and Amin, S. 2019. Knowledge distillation for end-to-end person search. *Proc. British Machine Vis. Conf.* .

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proc. Advances in Neural Inf. Process. Syst.*, 91–99.

Sun, Y.; Xu, Q.; Li, Y.; Zhang, C.; Li, Y.; Wang, S.; and Sun, J. 2019. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 393–402.

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 *Proc. Eur. Conf. Comp. Vis.*, 480–496.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive Representation Distillation. In *Proc. Int. Conf. Learn. Representations*.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2020. Fcos: A simple and strong anchor-free object detector. *IEEE Trans. Pattern Anal. Mach. Intell.* .

Tung, F.; and Mori, G. 2019. Similarity-preserving knowledge distillation. In *Proc. IEEE Int. Conf. Comp. Vis.*, 1365–1374.

Wang, C.; Ma, B.; Chang, H.; Shan, S.; and Chen, X. 2020a. TCTS: A Task-Consistent Two-Stage Framework for Person Search. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 11952–11961.

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. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 6449–6458.

Xiao, J.; Xie, Y.; Tillo, T.; Huang, K.; Wei, Y.; and Feng, J. 2019. IAN: the individual aggregation network for person search. *Pattern Recogn.* 87: 332–340.

Xiao, T.; Li, S.; Wang, B.; Lin, L.; and Wang, X. 2017. Joint detection and identification feature learning for person search. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 3415–3424.

Xu, Y.; Ma, B.; Huang, R.; and Lin, L. 2014. Person search in a scene by jointly modeling people commonness and person uniqueness. In *Proc. ACM Int. Conf. Multimedia*, 937–940.

Yan, Y.; Zhang, Q.; Ni, B.; Zhang, W.; Xu, M.; and Yang, X. 2019. Learning context graph for person search. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2158–2167.

Yuan, L.; Tay, F. E.; Li, G.; Wang, T.; and Feng, J. 2020. Revisiting Knowledge Distillation via Label Smoothing Regularization. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 3903–3911.

Zhang, L.; Shi, Z.; Zhou, J. T.; Cheng, M.-M.; Liu, Y.; Bian, J.-W.; Zeng, Z.; and Shen, C. 2020a. Ordered or Orderless: A Revisit for Video based Person Re-Identification. *IEEE Trans. Pattern Anal. Mach. Intell.* .

Zhang, X.; Cao, J.; Shen, C.; and You, M. 2019. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *Proc. IEEE Int. Conf. Comp. Vis.*, 8222–8231.

Zhang, X.; Gong, D.; Cao, J.; and Shen, C. 2020b. Memorizing comprehensively to learn adaptively: Unsupervised cross-domain person re-ID with multi-level memory. *arXiv preprint arXiv:2001.04123* .

Zhao, S.; Gao, C.; Zhang, J.; Cheng, H.; Han, C.; Jiang, X.; Guo, X.; Zheng, W.-S.; Sang, N.; and Sun, X. 2020. Do Not Disturb Me: Person Re-identification Under the Interference of Other Pedestrians. In *Proc. Eur. Conf. Comp. Vis.*, 647–663.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *Proc. IEEE Int. Conf. Comp. Vis.*, 1116–1124.

Zheng, L.; Zhang, H.; Sun, S.; Chandraker, M.; Yang, Y.; and Tian, Q. 2017. Person re-identification in the wild. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 1367–1376.

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proc. IEEE Int. Conf. Comp. Vis.*, 3754–3762.

Zhong, Y.; Wang, X.; and Zhang, S. 2020. Robust Partial Matching for Person Search in the Wild. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 6827–6835.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random Erasing Data Augmentation. In *Proc. AAAI Conf. Artificial Intell.*, 13001–13008.

Zhu, X.; Hu, H.; Lin, S.; and Dai, J. 2019. Deformable convnets v2: More deformable, better results. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 9308–9316.