Exploring Visual Context for Weakly Supervised Person Search

Yichao Yan[1*†], **Jinpeng Li**[2*], **Shengcai Liao**[2], **Jie Qin**[2], **Bingbing Ni**[1], **Ke Lu**[3], **Xiaokang Yang**[1]

[1] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China.
[2] Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates.
[3] University of Chinese Academy of Sciences, Beijing, China.
{yanyichao, nibingbing, xkyang}@sjtu.edu.cn, {ljpadam, qinjiebuaa}@gmail.com, scliao@ieee.org, luk@ucas.ac.cn

Abstract

Person search has recently emerged as a challenging task that jointly addresses pedestrian detection and person re-identification. Existing approaches follow a fully supervised setting where both bounding box and identity annotations are available. However, annotating identities is labor-intensive, limiting the practicability and scalability of current frameworks. This paper inventively considers weakly supervised person search with only bounding box annotations. We proposed to address this novel task by investigating three levels of context clues (i.e., detection, memory and scene) in unconstrained natural images. The first two are employed to promote local and global discriminative capabilities, while the latter enhances clustering accuracy. Despite its simple design, our CGPS achieves 80.0% in mAP on CUHK-SYSU, boosting the baseline model by 8.8%. Surprisingly, it even achieves comparable performance with several supervised person search models. Our code is available at https://github.com/ljpadam/CGPS

Introduction

Person search (Zheng et al. 2017; Xiao et al. 2017) aims to retrieve a query person from unconstrained natural images. It therefore needs to simultaneously address the tasks of pedestrian detection and person re-identification (re-id). Supervised person search in particular has been extensively studied in recent years (Liu et al. 2017; Xiao et al. 2019; Munjal et al. 2019). Nevertheless, annotating the identities remains labor-intensive. In contrast, relatively accurate bounding box annotations could be automatically generated from existing pedestrian detectors (Brazil, Yin, and Liu 2017; Liu et al. 2018, 2019). Motivated by this, in this work, we consider person search in the weakly supervised setting, in which there only exist bounding box annotations.

Intuitively, this task can be addressed with a two-step person search model, by first locating the pedestrians with a detector, and then applying an unsupervised person re-id model. However, there are two major issues with this straightforward solution. **1)** Two-step approaches employ two separate models to address pedestrian detection and person re-id, requiring twice as many parameters as one-step models, and introducing high computational overhead. Moreover, these models cannot achieve end-to-end inference and are thus inefficient. **2)** As existing unsupervised re-id methods take cropped person images as input, they do not have access to the rich context information when training the detection model, or the scene context. Therefore, directly extending an existing unsupervised person re-id model to address our task without considering the detection and scene context would make the solution suboptimal.

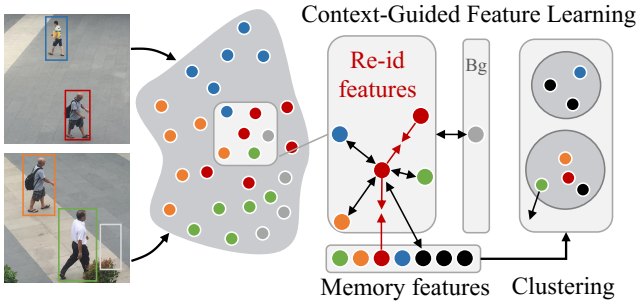


Figure 1: Overall pipeline of the proposed context-guided person search framework for weakly supervised person search. Without identity annotations, initial pseudo labels (colored points) are generated with ImageNet-pretrained weights. We employ the detection context to pull features belonging to the same identity together, while pushing the re-id features of the pedestrians away from the background features. A hard-negative mining strategy is designed to effectively employ the information in the memory. We use the scene context to generate more accurate clustering results.

To address the first issue, in this work, we propose a novel weakly supervised person search framework, termed as the Context-Guided Person Search Network (**CGPS**). Our model follows the typical architecture of a one-step person search method (Xiao et al. 2017), but it is only supervised with bounding box annotations. To learn re-id features, we adopt similar spirit of an unsupervised re-id model (Ge et al. 2020), and apply the clustering method to acquire human identities for unsupervised re-id training. This framework naturally inherits the advantages of existing one-step person search models in terms of efficiency, while at the same time avoiding the need for labor-intensive identity annotations.

*indicates equal contribution
†Corresponding author: Yichao Yan

In terms of the second issue, we further explore the context information for the joint learning framework to pursue discriminative representations for robust unsupervised re-id, as illustrated in Figure 1. Motivated by the recent advances in context learning (Doersch, Gupta, and Efros 2015; Zhang et al. 2018), the following visual context is investigated. **First**, we explore the *detection context* for re-id feature learning. Detection modules, such as Faster-RCNN (Ren et al. 2017), generate multiple positive predictions associated with each ground truth bounding box, as well as the negative predictions corresponding to the background. We observe that these samples can be naturally employed as a local constraint for learning identity features by keeping the re-id features belonging to the same bounding box close, while pushing the person and background features apart. **Second**, we investigate the global *memory context* to enable the model to pay more attention to hard negative samples in the feature memory. We design a sampling strategy based on the model's confidence on the negative samples, which adaptively selects the most effective hard negatives. **Third**, we employ the *scene context* to improve the clustering results. Previous methods naively cluster all the re-id features to generate pseudo-labels, without considering the natural context in the scene. We impose an intuitive constraint for clustering, i.e., different people appearing in an image should not belong to the same cluster. We demonstrate that this constraint significantly improves the clustering accuracy. These simple yet effective designs successfully transform an unsupervised person re-id model (Ge et al. 2020) into a promising weakly supervised person search framework, notably improving the searching performance.

In summary, our contributions are three-fold: (1) We propose a novel weakly supervised person search framework, which extends the current fully supervised paradigm. In addition, our weakly supervised task can be further extended to a fully unsupervised setting, requiring neither bounding box nor identity annotations on the target dataset. We expect the pioneering work to encourage future research in this direction. (2) We systematically investigate the visual context clues in the joint learning framework, for both feature learning and clustering. We demonstrate that these context clues work seamlessly in the framework to enhance the re-id feature learning. (3) As a weakly supervised person search framework, our model surprisingly outperforms several fully supervised models on CUHK-SYSU.

Related Work

Person Search. Person search is an extension of the person re-id task, which requires pedestrian detection and person re-id to be addressed in uncropped images. (Xu et al. 2014) first introduced this task, applying a sliding window search strategy with handcrafted features. However, it lacks proper benchmarks, until (Zheng et al. 2017) introduced a large dataset for person search, and systematically evaluated different combinations of several pedestrian detectors and re-id models. Meanwhile, (Xiao et al. 2017) proposed the first one-step detection and re-id network, enabling end-to-end learning. The follow-up works that have since been introduced can thus be grouped into one-step and two-step models. In general, two-step models (Chen et al. 2018; Lan, Zhu, and Gong 2018; Han et al. 2019; Wang et al. 2020) pay more attention to the consistency issue, i.e., how to learn discriminative re-id features based on the detection results. In contrast, one-step models (Liu et al. 2017; Chang et al. 2018; Dong et al. 2020a; Chen et al. 2020b) focus more on how to design an efficient person search framework.

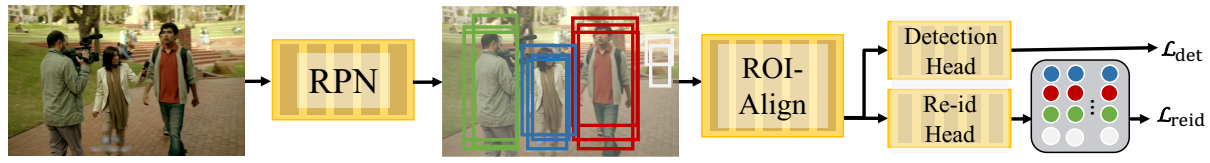
Despite their impressive progress, most of the current models are fully supervised. In this work, we explore the weakly supervised setting, i.e., we aim to learn the person search model with only bounding box annotations, while the re-id task is learned in an unsupervised manner.

Unsupervised Person Re-id. Early works, such as ELF (Gray and Tao 2008), LBP (Xiong et al. 2014) and LOMO (Liao et al. 2015), typically resort to handcrafted features. However, these models only achieve limited performance. Deep learning-based methods can generally be divided into two categories: (1) those that generate pseudo-labels for unsupervised instances (Lin et al. 2019; Zeng et al. 2020), and (2) those that translate labeled examples into the unlabeled domain (Song et al. 2019; Fu et al. 2019; Mekhazni et al. 2020). For the first category, pseudo-labels are typically generated by clustering the re-id embeddings. If videos are available, temporal constraints (Chen, Zhu, and Gong 2018; Li, Zhu, and Gong 2020) can be imposed to improve the clustering results. The second category is unsupervised domain adaptation, where the models are trained on the source domain and transferred to the unsupervised target domain. This can be achieved with generative adversarial networks (Wei et al. 2018; Deng et al. 2018; Liu, Chang, and Shen 2020; Chen, Zhu, and Gong 2019; Huang et al. 2019), clustering (Zhong et al. 2019; Fu et al. 2019; Zhai et al. 2020), and soft labels assigning (Wang and Zhang 2020).

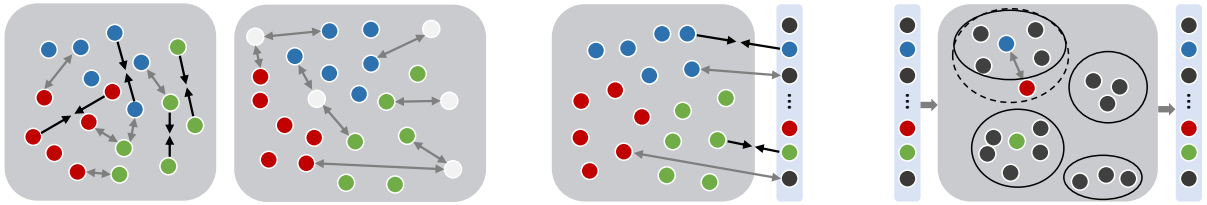
In this work, we consider the task without any re-id annotations. Therefore, we employ clustering to generate pseudo-labels for the re-id task. Furthermore, we extensively explore the context information in the joint learning framework to improve the quality of pseudo-labels and to enhance the discriminative ability of the re-id features.

Learning Visual Context. In addition to annotations, visual data implicitly contains rich context information, which has been widely explored in various vision tasks. For example, the relative location of object parts can be explicitly employed for visual representation learning (Doersch, Gupta, and Efros 2015; Noroozi and Favaro 2016). Meanwhile, the spatial and semantic relations between objects are helpful for the detection (Divvala et al. 2009; Dvornik, Mairal, and Schmid 2018; Barnea and Ben-Shahar 2019) and segmentation (Mottaghi et al. 2014; Zhang et al. 2018; Wang et al. 2021) tasks. In videos, the spatial-temporal relations can be modeled with a graph model to encode the interactions between objects, and enhance the performance of action recognition (Wang and Gupta 2018; Li et al. 2019). Some person recognition models (Huang, Xiong, and Lin 2018; Yan et al. 2019) have also recently employed the social/group context.

In our work, we simultaneously investigate three levels of context information, i.e., the detection context, the memory context, and the scene context, and further reveal their importance in weakly supervised person search.



(a) Pipeline of our framework



(b) Detection context (c) Memory context (d) Scene context

Figure 2: Illustration of our framework. (a) The basic architecture of our framework follows the typical person search model (Xiao et al. 2017), and we employ an unsupervised method (Ge et al. 2020) to train the re-id branch. (b) We employ the local context information generated by the detection task, to pull positive features belonging to the same instance together, and push background features apart. (c) We sample the hard negatives in the memory to facilitate global feature comparison. (d) Scene context is employed to generate more accurate clustering results.

Methodology

In this section, we first describe the overall framework for weakly supervised person search, and then delve into the details of the proposed context-guided learning strategies.

Framework Overview

Person Search Framework. In this work, we adopt the typical person search framework (Xiao et al. 2017) which jointly learns detection and re-id in an end-to-end manner. Specifically, this framework is developed upon the Faster-RCNN (Ren et al. 2017) architecture, as shown in Figure 2(a). Given an input image, the region proposal network (RPN) generates a set of pedestrian proposals, which are subsequently input into an RoI-Align layer to further extract the features of the candidate proposals. The output features are passed to a detection head and a re-id head, respectively. The detection head is trained with a classification loss and a regression loss, which supervise the person/non-person scores, as well as the bounding box locations. Meanwhile, the re-id head outputs the identity embedding associated with each pedestrian proposal, which is supervised with the pseudo-labels generated in an unsupervised manner.

Unsupervised Person Re-id. Among the recent methods (Lin et al. 2020, 2019; Zeng et al. 2020; Ge et al. 2020) designed for unsupervised re-id, we employ the memory-based loss (Ge et al. 2020) in our framework, for two reasons. First, it achieves strong performance on several unsupervised person re-id benchmarks, and is more likely to work well in our task. Second, as has been demonstrated in OIM (Xiao et al. 2017), memory-based learning schemes can effectively investigate the information in the unlabeled identities for person search. Specifically, the re-id features of all the training instances are stored in a memory $\boldsymbol{V} \in \mathbb{R}^{D\times N_a} = \{\boldsymbol{v}_1, ..., \boldsymbol{v}_{N_a}\}$, which contains N_a feature vectors with dimension D, where N_a is the number of instances in the training set. Through clustering, we obtain N_c clusters $\{\mathbb{C}_1, ..., \mathbb{C}_{N_c}\}$ with centroids $\boldsymbol{C} \in \mathbb{R}^{D\times N_c} = \{\boldsymbol{c}_1, ..., \boldsymbol{c}_{N_c}\}$. Note that we do not differentiate the clusters containing multiple instances from those with only one (i.e., the unclustered instance). Suppose the network outputs a re-id feature $\boldsymbol{x}_i$, the loss is defined as:

$$\mathcal{L}_i = -\log \frac{\exp(\boldsymbol{x}_i \cdot \boldsymbol{c}^+/\tau)}{\sum_{j=1}^{N_c} \exp(\boldsymbol{x}_i \cdot \boldsymbol{c}_j/\tau)}, \tag{1}$$

where $\boldsymbol{c}^+ = \boldsymbol{c}_j$ if $\boldsymbol{x}_i$ belongs to the j-th cluster, '$\cdot$' denotes the inner product, and $\tau > 0$ is a temperature hyper-parameter that controls the softness of the probability distribution. During backpropagation, the corresponding memory feature is updated by:

$$\boldsymbol{v} \leftarrow \gamma \boldsymbol{v} + (1-\gamma)\boldsymbol{x}_i, \tag{2}$$

where $\gamma \in [0, 1]$ controls the update ratio in the memory.

Context-Guided Feature Learning

We observe that the unsupervised re-id model trained with cropped instances does not make full use of the context information in person search. To better adapt the unsupervised re-id to person search, we investigate three types of context information, i.e., the detection context, the memory context, and the scene context.

Detection Context. When training the person search model, the RPN will generate several positive samples for each person, as well as some negative samples corresponding to the background. However, the loss function in Eq. 1 only employs the positive features, which yields two issues. 1) The input re-id features are only compared with memory features, while relations with the intra-batch re-id features are not explored. 2) There is no constraint on the background and foreground features. Although a few supervised person search models (Chen et al. 2020b,a) have addressed the second issue, the first issue tends to play a more important role in the unsupervised scenario. In this

work, we propose a quadruplet loss to simultaneously address both issues. Specifically, suppose the model outputs n_x re-id features $\boldsymbol{X} \in \mathbb{R}^{D \times n_x} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_{n_x}\}$ for the pedestrians, and n_b features corresponding to the backgrounds $\boldsymbol{B} \in \mathbb{R}^{D \times n_b} = \{\boldsymbol{b}_1, ..., \boldsymbol{b}_{n_b}\}$. For an instance $\boldsymbol{x}_i \in \boldsymbol{X}$, the loss function is defined as:

$$\begin{aligned}\mathcal{L}_i^{\mathrm{DC}} =& \alpha_1[-\min(\boldsymbol{x}_i \cdot \boldsymbol{x}^+) + \max(\boldsymbol{x}_i \cdot \boldsymbol{x}^-) + m]_+ \\ &+ \alpha_2[-\min(\boldsymbol{x}_i \cdot \boldsymbol{x}) + \max(\boldsymbol{x}_i \cdot \boldsymbol{b}) + m]_+,\end{aligned} \tag{3}$$

where $\boldsymbol{x}^+$ and $\boldsymbol{x}^-$ denote the positive and negative samples compared with $\boldsymbol{x}_i$, $\boldsymbol{x} \in \boldsymbol{X}$ and $\boldsymbol{b} \in \boldsymbol{B}$, while m denotes the distance margin, and α_1 and α_2 are the weights to balance the two loss terms. As illustrated in Figure 2(b), this function contains two terms. 1) An instance-to-instance term, which pulls the features belonging to the same instance close together, while pushing apart the features of different instances. 2) An instance-to-background term, which pushes the foreground features away from the background ones. In this way, we provide explicit instance-level guidance on the re-id feature, to yield more discriminative representations.

Memory Context. We observe that the negative samples contribute equally in Eq. 1. However, as indicated by prior works (Zeng et al. 2020; Chen et al. 2020a), putting more focus on the hard examples plays an important role in improving the model's discriminative capability. In this work, we explore the hard negative features in the memory, and propose a hard negative sampling strategy by evaluating the cumulative confidence of the hard-negative samples. Without loss of generality, all the negative features are sorted in descending order $\{\boldsymbol{c}_1^-, ..., \boldsymbol{c}_{N^-}^-\}$ according to their similarities with the input feature, where the number of negative samples is $N^- = N_c - 1$, indicating that the input feature only belongs to one specific cluster. We determine the number of hardest negative samples with:

$$K = \arg\min_k |\frac{\sum_{z=1}^{k} \boldsymbol{x}_i \cdot \boldsymbol{c}_z^-}{\sum_{j=1}^{N^-} \boldsymbol{x}_i \cdot \boldsymbol{c}_j^-} - \lambda|, \tag{4}$$

where λ is a threshold that controls the ratio of hard negative samples. In this case, Eq. 1 becomes:

$$\mathcal{L}_i^{\mathrm{MC}} = -\log \frac{\exp(\boldsymbol{x}_i \cdot \boldsymbol{c}^+/\tau)}{\exp(\boldsymbol{x}_i \cdot \boldsymbol{c}^+/\tau) + \sum_{k=1}^{K} \exp(\boldsymbol{x}_i \cdot \boldsymbol{c}_k^-/\tau)}. \tag{5}$$

Compared to sampling a fixed size of hard negative samples, our sampling strategy is more flexible to focus on a small set of the most effective negative samples.

The overall loss function is a combination of the targets in Eq. 3 and Eq. 5:

$$\mathcal{L}_{\mathrm{reid}} = \sum_i (\mathcal{L}_i^{\mathrm{DC}} + \mathcal{L}_i^{\mathrm{MC}}). \tag{6}$$

We find these two terms are complementary to each other: $\mathcal{L}_i^{\mathrm{DC}}$ tries to discriminate the local features within a batch, while $\mathcal{L}_i^{\mathrm{MC}}$ focuses more on the global discrimination by comparing the output feature with all the instances in the memory. Therefore, combining them is likely to yield more discriminative representations.

Scene Context. To assign pseudo-labels for the training instances, existing works (Ge et al. 2020; Chen et al. 2020c) directly perform clustering (e.g., DBSCAN (Ester et al. 1996)) on the instance-level re-id features. However, the results of the unsupervised clustering are not perfect, which directly influences the re-id feature learning in the framework. Although some works have tried to improve the clustering reliability with self-paced learning (Ge et al. 2020) or instance discrimination learning (Chen et al. 2020c), these works only consider regularizing the instance-level representation, neglecting the relative relations between examples in the scene. In this work, we impose a simple constraint on the clustering results: *the persons appearing in the same image cannot belong to the same cluster*. Specifically, after generating the clustering results based on the memory features, we further go through each cluster and split the clusters containing multiple instances belonging to the same image. For a certain image, only a single instance that is closest to the cluster center will be retained, while other instances will be removed and become unclustered instances. As shown in Figure 2 (d), the original top left cluster contains the blue and red instances. However, since they belong to the same image, the red one, which is far away from the center, is excluded from the cluster. In this way, we are able to generate better clustering results without any additional annotation. Suppose we have N_I training images, and I_i contains the instances in the i-th image. Give $\boldsymbol{I} = \{I_1, ..., I_{N_I}\}$ and the memory feature $\boldsymbol{V}$, we summarize the clustering with scene context in Algorithm 1.

Algorithm 1: Clustering with Scene Context

Require: $\boldsymbol{I}, \boldsymbol{V}$
Cluster $\boldsymbol{V}$ with DBSCAN $\rightarrow \mathbb{C}_1, ..., \mathbb{C}_{N_c}$
for $i \leftarrow 1, ..., N_c$ **do** ▷*iterate each cluster*
  for $j \leftarrow 1, ..., N_I$ **do** ▷*iterate each image*
    $\{\boldsymbol{v}_{i,j}^1, ..., \boldsymbol{v}_{i,j}^K\} \leftarrow \mathbb{C}_i \cap I_j$
      ▷*find the instances belonging to $\mathbb{C}_i$ and I_j*
    $l \leftarrow \arg\max_k (\boldsymbol{v}_{i,j}^k \cdot \boldsymbol{c}_i)$
      ▷*find the instance nearest to the cluster center*
    for $k \leftarrow 1, ..., l-1, l+1, ..., K$ **do**
      $\mathbb{C}_i \leftarrow \mathbb{C}_i - \boldsymbol{v}_{i,j}^k$ ▷*remove this instance from $\mathbb{C}_i$*
      $N_c \leftarrow N_c + 1$ ▷*update the number of clusters*
      $\mathbb{C}_{N_c} \leftarrow \{\boldsymbol{v}_{i,j}^k\}$ ▷*add a new cluster*
    end for
  end for
end for

Discussions

Although some similar spirits have been adopted in prior supervised models, our work is essentially different from them. *First*, TCTS (Wang et al. 2020) and AlignPS (Yan et al. 2021) employ the within-image relation in the training loss, while we consider this as a constraint for generating more accurate pseudo labels. *Second*, although several supervised models (Chen et al. 2020a,b) consider pushing the person and background features apart, our quadruplet loss additionally explores the relations of intra-batch re-id features. *Third*, we investigate three context clues to explicitly tackle the task of weakly supervised person search, while existing model focus on fully supervised setting.

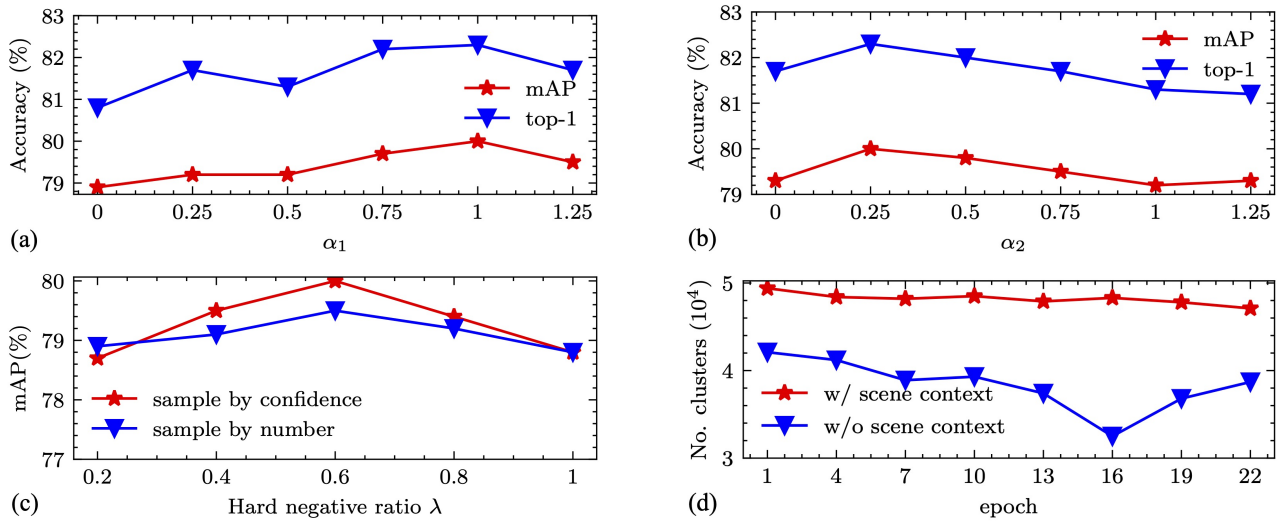


Figure 3: Visualization of the contribution of different types of context clues on CUHK-SYSU. (a) The contribution of the instance-to-instance term in the detection context. (b) The impact of the instance-to-background term in the detection context. (c) The influence of hard negative sampling. (d) The contribution of scene context.

Experiments

Experimental Setup

Datasets. We evaluate the proposed framework on the following two datasets. **CUHK-SYSU** (Xiao et al. 2017) is one of the largest public datasets designed for person search, which contains 18,184 images captured from streets and TV/movie frames. This dataset also includes 96,143 bounding box annotations, with 8,432 different identities. **PRW** (Zheng et al. 2017) was collected from six surveillance cameras. It contains 11,816 video frames, with 43,110 annotated bounding boxes and 932 identities.

Evaluation Protocol. We employ the standard `train/test` splits for both CUHK-SYSU and PRW. Only the bounding box annotations are employed in the training phase, i.e., 11,216 images with 55,272 bounding boxes for CUHK-SYSU, and 5,705 images with 18,048 boxes for PRW. The test set of CUHK-SYSU contains 2,900 query persons and 6,978 images. In the following sections, we report the results with 100 gallery images unless otherwise specified. PRW contains a test set with 2,057 queries and 6,112 images. We report the mean average precision (mAP) and top-1 ranking accuracy as evaluation metrics.

Implementation Details This project is supported by MindSpore [1]. Following (Yan et al. 2021), we employ a multi-scale training strategy, while input images are resized to a fixed size of 1500×900 for inference. In the training phase, random flipping is applied and we employ stochastic gradient descent (SGD) optimizer, with the weight decay set to 0.0005. We set the batch size to 4 and initialize the learning rate to 0.0012, which is reduced by a factor of 10 at epoch 16, training to a total of 22 epochs. We set the default hyper-parameters $\gamma = 0.2$, $\alpha_1 = 1$, $\alpha_2 = 0.25$, and $\lambda = 0.6$. We employ DBSCAN (Ester et al. 1996) with self-paced training (Ge et al. 2020) as the clustering method. We set $\epsilon = 0.7$, while other hyperparameters follow SPCL (Ge et al. 2020).

[1]https://www.mindspore.cn/

Detection context	Memory context	Scene context	CUHK-SYSU mAP	CUHK-SYSU top-1
			71.2	73.8
√			75.5 (+4.3)	78.6 (+4.8)
	√		71.3 (+0.1)	74.1 (+0.3)
		√	77.1 (+5.9)	79.9 (+6.1)
√	√		75.8 (+4.6)	78.8 (+5.0)
√		√	78.8 (+7.6)	80.9 (+7.1)
	√	√	78.7 (+7.5)	81.2 (+7.4)
√	√	√	**80.0** (**+8.8**)	**82.3** (**+8.5**)

Table 1: Comparative results on CUHK-SYSU when employing different context learning strategies.

Analytical Results

Comparative Results. We first evaluate the effectiveness of the proposed context-guided learning strategies. We compare the baseline method with different combinations of context information, and report the results on CUHK-SYSU in Table 1. We observe that the baseline framework achieves 71.2% mAP and 73.8% top-1 accuracy without any identity annotation. With the help of detection, memory and scene context, the model achieves 4.3%, 0.1% and 5.9% improvements in mAP, respectively. Furthermore, by combining these contexts, the model's performance is further consistently improved. Finally, the model achieves 80.0% mAP when employing all the context, significantly improving the baseline model by **8.8**% in mAP.

Detection Context. We visualize the contributions of the instance-to-instance term and instance-to-background term in Figure 3(a) and 3(b), by varying α_1 and α_2 in Eq. 3. We observe that both terms play positive roles in re-id feature learning, while the best performance is achieved with $\alpha_1 = 1$ and $\alpha_2 = 0.25$. To differentiate the re-id features of pedestrians from the background, we employ the instance-to-background term to directly push the two types of features apart, which proves to be effective in our framework.

	Methods	CUHK		PRW	
		mAP	R1	mAP	R1
fully supervised	OIM (Xiao et al. 2017)	75.5	78.7	21.3	49.4
	IDE (Zheng et al. 2017)	-	-	20.5	48.3
	IAN (Xiao et al. 2019)	76.3	80.1	23.0	61.9
	NPSM (Liu et al. 2017)	77.9	81.2	24.2	53.1
	RCAA(Chang et al. 2018)	79.3	81.3	-	-
	MGTS (Chen et al. 2018)	83.0	83.7	32.6	72.1
	CTXG (Yan et al. 2019)	84.1	86.5	33.4	73.6
	BINet (Dong et al. 2020a)	90.0	90.7	45.3	81.7
	NAE+ (Chen et al. 2020b)	92.1	92.9	44.0	81.1
	IGPN (Dong et al. 2020b)	90.3	91.4	47.2	87.0
	RDLR (Han et al. 2019)	93.0	94.2	42.9	70.2
	TCTS (Wang et al. 2020)	93.9	95.1	46.8	87.5
	AlignPS (Yan et al. 2021)	93.1	93.4	45.9	81.9
	SeqNet(Li and Miao 2021)	**94.8**	**95.7**	**47.6**	**87.6**
Ours (weakly sup)		**80.0**	**82.3**	**16.2**	**68.0**

Table 2: Comparison with the state-of-the-art supervised person search models.

Memory Context. To evaluate the effectiveness of the memory context, we compare it with the other widely employed "sample by number" strategy, i.e., selecting the top λN^- hardest negatives from all the negative samples. As can be observed from Figure 3(c), our model is sensitive to this hyperparameter. We observe that when $\lambda < 0.4$, the performance significantly decreases as too many easy samples are ignored. However, by selecting a proper hard negative ratio (i.e., $0.4 \sim 0.8$), our proposed strategy achieves notable improvements compared with the "sample by number" strategy, as well as the baseline method without hard negative sampling ($\lambda = 1$). These results validate the effectiveness of the hard negatives sampled by our strategy.

Scene Context. To reveal how the scene context impacts the clustering results, we visualize the evolution of cluster numbers during training. As shown in Figure 3(d), the number of clusters decreases smoothly, indicating that more and more samples are clustered together. However, there exists a significant gap with and without the scene context constraint, i.e., a large number of (about 10,000) samples are incorrectly clustered at each epoch. We employ the scene context to remove these incompatible samples from the clusters, hence improving the quality of the pseudo-labels. As scene context directly affects the pseudo-labels, we notice that it contributes the most among all the three context clues, demonstrating its importance.

Comparison with Supervised Models

To further evaluate the performance of our weakly supervised framework, we compare it with the fully supervised models. Surprisingly, our weakly supervised framework outperforms several supervised models on CUHK-SYSU, e.g., OIM (Xiao et al. 2017), IAN (Xiao et al. 2019) and NPSM (Liu et al. 2017). Compared with the state-of-the-art supervised models, e.g., SeqNet (Li and Miao 2021), NAE (Chen et al. 2020b), BINet (Dong et al. 2020a), there still exists a significant margin in performance. We therefore hope this work will provide a starting point to enable future works to explore solutions for bridging this gap.

Methods	GFLOPS	Time	mAP	top1
FRCNN + SBL	346	101	67.2	68.9
FRCNN + SPCL	346	100	71.8	72.1
FRCNN + MMT	346	101	73.4	74.9
Ours	**281**	**68**	**80.0**	**82.3**

Table 3: Comparison with two-step models on CUHK-SYSU, w.r.t. performance and efficiency. SBL denotes the strong baseline in (Ge et al. 2020). Runtime is measured by the average inference time in millisecond (ms).

Methods	GFLOPS	Time	mAP	top1
ResNet-18	**146**	**31**	60.1	63.2
ResNet-34	235	40	67.5	70.8
ResNet-50	281	68	80.0	82.3
ResNet-101	327	81	**80.4**	**82.5**

Table 4: Results on CUHK-SYSU with various backbones.

Comparison with Two-Step Models

To evaluate the performance and efficiency of the proposed framework, we compare it with several two-step models, which first localize pedestrians with a detector (Faster-RCNN), and then employ an unsupervised re-id method (Ge et al. 2020; Ge, Chen, and Li 2020) for person retrieval. As can be seen from Table 3, our one-step method not only outperforms the two-step models, but also displays significant advantage in terms of efficiency. Specifically, although our model is based on the re-id training prototype SPCL (Ge et al. 2020), the context information helps it outperform its two-step counterpart by a large margin. For the model complexity and runtime analysis, we employ the same backbone (i.e., ResNet-50) for all the two-step models. Therefore, they obtain similar FLOPS and runtime during inference. As our model only needs a single forward pass to generate both detection results and re-id features, it displays lower FLOPS (346G $\rightarrow$ 281G) and shorter runtime (101 ms $\rightarrow$ 68 ms). It is also noteworthy that the parameters of the two-step models are twice as our framework, as they adopt two separate ResNet-50 models for detection and re-id, respectively.

Further Discussions

Different Backbones. As shown in Table 4, using different backbones considerably affects our model's performance. With ResNet-18, the model achieves 60.1% in mAP, while its performance is significantly improved to 80.4% when employing ResNet-101. However, deeper backbone networks also bring significantly higher FLOPS (146G $\rightarrow$ 327G) and longer runtime (31 ms $\rightarrow$ 81 ms). In practice, a suitable backbone needs to be carefully selected to meet the requirements of real-world applications.

Training Samples. In Figure 5, we illustrate the impact of employing different numbers of training samples on CUHK-SYSU. We observe that more training samples generate better results, which motivates us to employ external training data. Following this idea, we combine CUHK-SYSU, PRW, and an external dataset COCO (Lin et al. 2014) for training. Specifically, we select 6,500 images containing pedestrians

(a) Query (b) CGPS w/o context (c) CGPS (a) Query (b) CGPS w/o context (c) CGPS

Figure 4: Difficult cases that can be successfully retrieved by CGPS, but not CGPS without context information. The yellow bounding boxes denote the queries, while the green and red ones denote correct and incorrect top-1 matches, respectively.

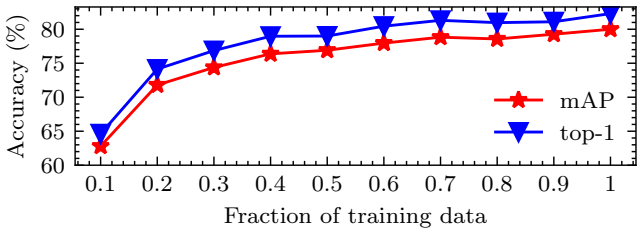


Figure 5: Results on CUHK-SYSU with different numbers of samples for weakly supervised training.

Training Data			CUHK-SYSU		PRW	
CUHK	PRW	COCO	mAP	top-1	mAP	top-1
√			80.0	82.3	15.0	67.6
	√		60.5	63.1	16.2	68.0
√	√		**80.5**	**82.8**	**18.7**	**73.2**
√		√	74.8	77.0	13.0	66.5
	√	√	71.7	74.3	16.4	69.7
√	√	√	75.8	78.6	15.7	70.8

Table 5: Comparative results on CUHK-SYSU and PRW when employing different combinations of training data.

from the original COCO dataset, and only utilize pedestrian bounding box annotations. As shown in Table 5, the combination of CUHK-SYSU and PRW achieves the best performance on both datasets. Furthermore, the improvement on PRW is more significant, maybe due to the fact that PRW contains fewer training samples. However, adding more training data does not always bring improvement, e.g., the combination of CUHK-SYSU and COCO achieves inferior performance compared with employing CUHK-SYSU alone. This is because of the domain gaps between these datasets; therefore, considering domain adaptation is another direction worth exploration in the future.

Extension to Unsupervised Person Search. Our weakly supervised framework can be naturally extended to a fully unsupervised setting by generating the bounding box annotations from existing pedestrian detectors, e.g, Cascade R-CNN (Cai and Vasconcelos 2018), RetinaNet (Lin et al. 2017), and EMD+RM (Chu et al. 2020). We compare the model trained with the ground-truth (GT) bounding boxes and the predicted ones from these pedestrian detectors trained on *CrowdHuman* (Shao et al. 2018). The comparison results are reported in Table 6. We observe that no matter which detector is employed, we can obtain satisfactory performance on both datasets, which is only slightly lower compared with the models trained with GT bounding boxes. This shows the potential of building a person search framework completely free of manual annotations.

Annotation	CUHK-SYSU		PRW	
	mAP	top-1	mAP	top-1
Cascade R-CNN	76.2	79.2	14.7	66.5
RetinaNet	76.0	78.5	15.1	65.8
EMD+RM	76.6	80.1	14.6	66.6
GT	80.0	82.3	16.2	68.0

Table 6: Comparative results on CUHK-SYSU and PRW when employing different bounding box annotations.

Qualitative Results. We visualize some qualitative results in Figure 4. As can be observed, our model successfully retrieves the target person in several challenging cases, where the baseline model without the context information fails. As can be seen from the top-left example, CGPS retrieves the correct target from a different viewpoint. In contrast, the baseline model returns another person with similar appearance. Other examples show that our context-guided model is more robust to the situations of occlusion, as well as illumination/scale variations. These results demonstrate the importance of the context clues in our framework.

Conclusion

This paper investigates weakly supervised person search, a novel task with only bounding box locations, avoiding the need to collect labor-intensive identity annotations. By extensively exploring the detection, memory, and scene context, we successfully develop a weakly supervised person search framework. Our framework achieves promising results on two benchmarks, but there still exists a gap compared with supervised models. We expect that this work will foster future research towards solving this.

Acknowledgements

This work was supported by CAAI-Huawei MindSpore Open Fund, Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102).

References

Barnea, E.; and Ben-Shahar, O. 2019. Exploring the Bounds of the Utility of Context for Object Detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 7412–7420.

Brazil, G.; Yin, X.; and Liu, X. 2017. Illuminating Pedestrians via Simultaneous Detection and Segmentation. In *Int. Conf. Comput. Vis.*, 4960–4969.

Cai, Z.; and Vasconcelos, N. 2018. Cascade R-CNN: Delving Into High Quality Object Detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 6154–6162.

Chang, X.; Huang, P.; Shen, Y.; Liang, X.; Yang, Y.; and Hauptmann, A. G. 2018. RCAA: Relational Context-Aware Agents for Person Search. In *Eur. Conf. Comput. Vis.*, 86–102.

Chen, D.; Zhang, S.; Ouyang, W.; Yang, J.; and Schiele, B. 2020a. Hierarchical Online Instance Matching for Person Search. In *AAAI*, 10518–10525.

Chen, D.; Zhang, S.; Ouyang, W.; Yang, J.; and Tai, Y. 2018. Person Search via a Mask-Guided Two-Stream CNN Model. In *Eur. Conf. Comput. Vis.*, 764–781.

Chen, D.; Zhang, S.; Yang, J.; and Schiele, B. 2020b. Norm-Aware Embedding for Efficient Person Search. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 12612–12621.

Chen, G.; Lu, Y.; Lu, J.; and Zhou, J. 2020c. Deep Credible Metric Learning for Unsupervised Domain Adaptation Person Re-identification. In *Eur. Conf. Comput. Vis.*, 643–659.

Chen, Y.; Zhu, X.; and Gong, S. 2018. Deep Association Learning for Unsupervised Video Person Re-identification. In *Brit. Mach. Vis. Conf.*, 48.

Chen, Y.; Zhu, X.; and Gong, S. 2019. Instance-Guided Context Rendering for Cross-Domain Person Re-Identification. In *Int. Conf. Comput. Vis.*, 232–242.

Chu, X.; Zheng, A.; Zhang, X.; and Sun, J. 2020. Detection in Crowded Scenes: One Proposal, Multiple Predictions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 12211–12220.

Deng, W.; Zheng, L.; Ye, Q.; Kang, G.; Yang, Y.; and Jiao, J. 2018. Image-Image Domain Adaptation With Preserved Self-Similarity and Domain-Dissimilarity for Person Re-Identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 994–1003.

Divvala, S. K.; Hoiem, D.; Hays, J.; Efros, A. A.; and Hebert, M. 2009. An empirical study of context in object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 1271–1278.

Doersch, C.; Gupta, A.; and Efros, A. A. 2015. Unsupervised Visual Representation Learning by Context Prediction. In *Int. Conf. Comput. Vis.*, 1422–1430.

Dong, W.; Zhang, Z.; Song, C.; and Tan, T. 2020a. Bi-Directional Interaction Network for Person Search. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2836–2845.

Dong, W.; Zhang, Z.; Song, C.; and Tan, T. 2020b. Instance Guided Proposal Network for Person Search. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2582–2591.

Dvornik, N.; Mairal, J.; and Schmid, C. 2018. Modeling Visual Context Is Key to Augmenting Object Detection Datasets. In *Eur. Conf. Comput. Vis.*, 375–391.

Ester, M.; Kriegel, H.; Sander, J.; and Xu, X. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *International Conference on Knowledge Discovery and Data Mining*, 226–231.

Fu, Y.; Wei, Y.; Wang, G.; Zhou, Y.; Shi, H.; and Huang, T. S. 2019. Self-Similarity Grouping: A Simple Unsupervised Cross Domain Adaptation Approach for Person Re-Identification. In *Int. Conf. Comput. Vis.*, 6111–6120.

Ge, Y.; Chen, D.; and Li, H. 2020. Mutual Mean-Teaching: Pseudo Label Refinery for Unsupervised Domain Adaptation on Person Re-identification. In *International Conference on Learning Representations*.

Ge, Y.; Zhu, F.; Chen, D.; Zhao, R.; and Li, H. 2020. Self-paced Contrastive Learning with Hybrid Memory for Domain Adaptive Object Re-ID. In *Adv. Neural Inform. Process. Syst.*

Gray, D.; and Tao, H. 2008. Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In *Eur. Conf. Comput. Vis.*, 262–275.

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 *Int. Conf. Comput. Vis.*, 9813–9822.

Huang, Q.; Xiong, Y.; and Lin, D. 2018. Unifying Identification and Context Learning for Person Recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2217–2225.

Huang, Y.; Wu, Q.; Xu, J.; and Zhong, Y. 2019. SBSGAN: Suppression of Inter-Domain Background Shift for Person Re-Identification. In *Int. Conf. Comput. Vis.*, 9526–9535.

Lan, X.; Zhu, X.; and Gong, S. 2018. Person Search by Multi-Scale Matching. In *Eur. Conf. Comput. Vis.*, volume 11205, 553–569.

Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; and Tian, Q. 2019. Actional-Structural Graph Convolutional Networks for Skeleton-Based Action Recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 3595–3603.

Li, M.; Zhu, X.; and Gong, S. 2020. Unsupervised Tracklet Person Re-Identification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(7): 1770–1782.

Li, Z.; and Miao, D. 2021. Sequential End-to-end Network for Efficient Person Search. In *AAAI*, 2011–2019.

Liao, S.; Hu, Y.; Zhu, X.; and Li, S. Z. 2015. Person re-identification by Local Maximal Occurrence representation and metric learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2197–2206.

Lin, T.; Goyal, P.; Girshick, R. B.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. In *Int. Conf. Comput. Vis.*, 2999–3007.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft

COCO: Common Objects in Context. In *Eur. Conf. Comput. Vis.*, 740–755.

Lin, Y.; Dong, X.; Zheng, L.; Yan, Y.; and Yang, Y. 2019. A Bottom-Up Clustering Approach to Unsupervised Person Re-Identification. In *AAAI*, 8738–8745.

Lin, Y.; Xie, L.; Wu, Y.; Yan, C.; and Tian, Q. 2020. Unsupervised Person Re-Identification via Softened Similarity Learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 3387–3396.

Liu, C.; Chang, X.; and Shen, Y. 2020. Unity Style Transfer for Person Re-Identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 6886–6895.

Liu, H.; Feng, J.; Jie, Z.; Karlekar, J.; Zhao, B.; Qi, M.; Jiang, J.; and Yan, S. 2017. Neural Person Search Machines. In *Int. Conf. Comput. Vis.*, 493–501.

Liu, W.; Liao, S.; Hu, W.; Liang, X.; and Chen, X. 2018. Learning Efficient Single-Stage Pedestrian Detectors by Asymptotic Localization Fitting. In *Eur. Conf. Comput. Vis.*, 643–659.

Liu, W.; Liao, S.; Ren, W.; Hu, W.; and Yu, Y. 2019. High-Level Semantic Feature Detection: A New Perspective for Pedestrian Detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 5187–5196.

Mekhazni, D.; Bhuiyan, A.; Ekladious, G. S. E.; and Granger, E. 2020. Unsupervised Domain Adaptation in the Dissimilarity Space for Person Re-identification. In *Eur. Conf. Comput. Vis.*, 159–174.

Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.; Lee, S.; Fidler, S.; Urtasun, R.; and Yuille, A. L. 2014. The Role of Context for Object Detection and Semantic Segmentation in the Wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 891–898.

Munjal, B.; Amin, S.; Tombari, F.; and Galasso, F. 2019. Query-Guided End-To-End Person Search. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 811–820.

Noroozi, M.; and Favaro, P. 2016. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In *Eur. Conf. Comput. Vis.*, 69–84.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1137–1149.

Shao, S.; Zhao, Z.; Li, B.; Xiao, T.; Yu, G.; Zhang, X.; and Sun, J. 2018. CrowdHuman: A Benchmark for Detecting Human in a Crowd. *CoRR*, abs/1805.00123.

Song, J.; Yang, Y.; Song, Y.; Xiang, T.; and Hospedales, T. M. 2019. Generalizable Person Re-Identification by Domain-Invariant Mapping Network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 719–728.

Wang, C.; Ma, B.; Chang, H.; Shan, S.; and Chen, X. 2020. TCTS: A Task-Consistent Two-Stage Framework for Person Search. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 11949–11958.

Wang, D.; and Zhang, S. 2020. Unsupervised Person Re-Identification via Multi-Label Classification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 10978–10987.

Wang, W.; Zhou, T.; Yu, F.; Dai, J.; Konukoglu, E.; and Van Gool, L. 2021. Exploring Cross-Image Pixel Contrast for Semantic Segmentation. In *Int. Conf. Comput. Vis.*, 7303–7313.

Wang, X.; and Gupta, A. 2018. Videos as Space-Time Region Graphs. In *Eur. Conf. Comput. Vis.*, 413–431.

Wei, L.; Zhang, S.; Gao, W.; and Tian, Q. 2018. Person Transfer GAN to Bridge Domain Gap for Person Re-Identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 79–88.

Xiao, J.; Xie, Y.; Tillo, T.; Huang, K.; Wei, Y.; and Feng, J. 2019. IAN: The Individual Aggregation Network for Person Search. *Pattern Recognit.*, 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 3376–3385.

Xiong, F.; Gou, M.; Camps, O. I.; and Sznaier, M. 2014. Person Re-Identification Using Kernel-Based Metric Learning Methods. In *Eur. Conf. Comput. Vis.*, 1–16.

Xu, Y.; Ma, B.; Huang, R.; and Lin, L. 2014. Person Search in a Scene by Jointly Modeling People Commonness and Person Uniqueness. In *ACM Int. Conf. Multimedia*, 937–940.

Yan, Y.; Li, J.; Qin, J.; Bai, S.; Liao, S.; Liu, L.; Zhu, F.; and Shao, L. 2021. Anchor-Free Person Search. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 7690–7699.

Yan, Y.; Zhang, Q.; Ni, B.; Zhang, W.; Xu, M.; and Yang, X. 2019. Learning Context Graph for Person Search. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2158–2167.

Zeng, K.; Ning, M.; Wang, Y.; and Guo, Y. 2020. Hierarchical Clustering With Hard-Batch Triplet Loss for Person Re-Identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 13654–13662.

Zhai, Y.; Lu, S.; Ye, Q.; Shan, X.; Chen, J.; Ji, R.; and Tian, Y. 2020. AD-Cluster: Augmented Discriminative Clustering for Domain Adaptive Person Re-Identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 9018–9027.

Zhang, H.; Dana, K. J.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; and Agrawal, A. 2018. Context Encoding for Semantic Segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 7151–7160.

Zheng, L.; Zhang, H.; Sun, S.; Chandraker, M.; Yang, Y.; and Tian, Q. 2017. Person Re-identification in the Wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 3346–3355.

Zhong, Z.; Zheng, L.; Luo, Z.; Li, S.; and Yang, Y. 2019. Invariance Matters: Exemplar Memory for Domain Adaptive Person Re-Identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 598–607.