Context-Aware Feature Learning for Noise Robust Person Search

Cairong Zhao, Zhicheng Chen, Shuguang Dou, Zefan Qu, Jiawei Yao, Jun Wu, *Senior Member, IEEE*, and Duoqian Miao

***Abstract*— Person search aims to localize and identify specific pedestrians from numerous surveillance scene images. In this work, we focus on the noise in person search. We categorize the noise into *scene-inherent* noise and *human-introduced* noise. Scene-inherent noise comes from congestion, occlusion, and illumination changes. Human-introduced noise originates from the labeling process. For *scene-inherent* noise, we propose a novel context contrastive loss to take advantage of the latent contextual information from scene images. Features from context regions are utilized to construct contrastive pairs to constrain the feature discrimination among pedestrians in scene images while maintaining the feature consistency of the same identity. The network can thus learn to distinguish congested and overlapped pedestrians and more robust features can be obtained. For *human-introduced* noise, we propose a noise-discovery and noise-suppression training process for mislabeling robust person search. After the first training pass, the relation between feature prototypes of different identities is analyzed and the mislabeled pedestrians are discovered. During the second training pass, the label noise is suppressed to reduce the negative influence of mislabeled data. Experiments show that the proposed context-aware noise-robust (CANR) person search can achieve competitive performance. Further ablation studies confirm the effectiveness of CANR.**

***Index Terms*— Person search, person re-identification, pedestrian detection, contrastive learning.**

I. Introduction

With the increasing demand for security, numerous surveillance videos are being captured. Processing and making good use of them is a challenging task. Person re-identification [1], [2] is a research area trying to resolve this challenge. However, the problem setting of person re-identification requires cropped pedestrian images. So, there is a gap between the research of person re-identification and application in real-world scenarios. Simply cascading pedestrian detection [3]–[5] and person re-identification is sub-optimal. Errors and inaccurate results from upstream object detection will damage the performance of person re-identification. Person search [6]–[8] is an emerging research area, which integrates pedestrian detection and person re-identification into a unified framework. As a result, optimizing both parts simultaneously will lead to better performance and is closer to real-world applications.

Existing research on person search can be divided into two categories: *End-to-end methods* and *two-step methods*. End-to-end methods integrate pedestrian detection and person re-identification into a single network where both two sub-tasks of person search are optimized simultaneously [7], [9]–[14]. Two-step methods start with two cascaded sub-tasks and optimize each sub-task for mutual adaption so that they can be better integrated [15]–[19]. Within those two categories, a number of methods explored similar ideas like leveraging query identity as guidance for feature extraction [12], [17], [18], using extra supervision from masks [14], [19], key-points [13], [14], contextual information [20]–[22]. However, only a few works had ever reflected on the fundamental difference between person search and a detection-identification pipeline. Most of the existing person search methods ignore the spatial relationship of pedestrians and the latent contextual information that comes with scene images. Some works [21], [22] utilized the contextual information with the assumption that the co-traveler of the target person will be relatively consistent. The assumption of consistent to-travelers does hold due to limitations of the existing datasets: *The same pedestrians in the datasets are collected from several camera views of the same or nearby geographical location.* Their performance under more realistic scenarios remains unknown.

In this work, we focus on the noise in person search and take a different approach to utilize the context information from the scene images. We categorize the noise in person search into *scene-inherent* noise and *human-introduced* noise. *Scene-inherent* noise comes from congestion, occlusion, illumination variation, etc while *human-introduced* noise comes from mislabeled data during labeling. Examples are shown in Fig.2(a) and Fig.2(b) respectively. Each image in Fig.2(a)

Manuscript received 16 March 2022; revised 2 May 2022 and 23 May 2022; accepted 28 May 2022. Date of publication 30 May 2022; date of current version 4 October 2022. This work was supported in part by the National Natural Science Fund of China under Grant 62076184, Grant 61673299, Grant 61976160, and Grant 62076182; in part by the Shanghai Innovation Action Project of Science and Technology under Grant 20511100700; in part by the Shanghai Natural Science Foundation of Shanghai under Grant 22ZR1466700; in part by the Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0100; and in part by the Fundamental Research Funds for the Central Universities. This article was recommended by Associate Editor V. Stankovic. *(Cairong Zhao and Zhicheng Chen contributed equally to this work.) (Corresponding author: Cairong Zhao.)*

Cairong Zhao, Zhicheng Chen, Shuguang Dou, Zefan Qu, and Duoqian Miao are with the Department of Computer Science and Technology, Tongji University, Shanghai 201804, China (e-mail: zhaocairong@tongji.edu.cn).

Jiawei Yao is with College of Architecture and Urban Planning, Tongji University, Shanghai 200092, China (e-mail: jiawei.yao@tongji.edu.cn).

Jun Wu is with the School of Computer Science, Fudan University, Shanghai 200438, China.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2022.3179441.

Digital Object Identifier 10.1109/TCSVT.2022.3179441

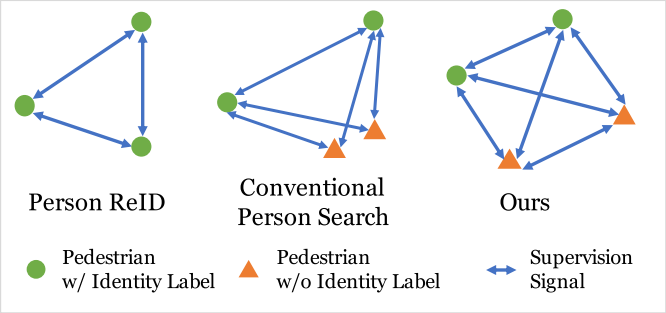


Fig. 1. Illustration of the difference in feature level supervision between different approaches. Our work includes unlabeled pedestrians in the training process and fully utilizes the latent context information.

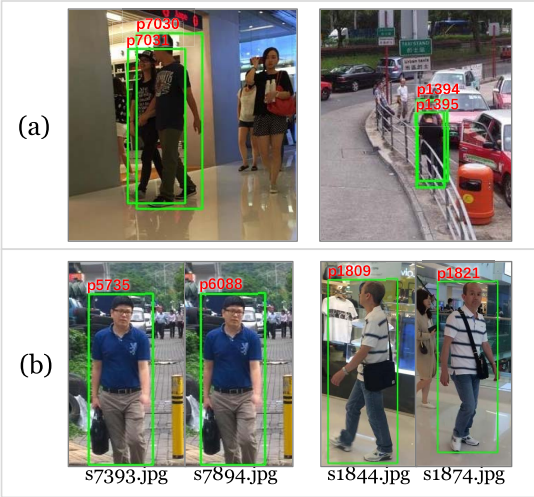


Fig. 2. Examples of (a) scene-inherent noise and (b) human-introduced noise from the CUHK-SYSU dataset. The '*.jpg' below demonstrates the filename of the image in the dataset. The 'p*' in red on the top left corner of each bounding box denotes the identity label of each pedestrian.

shows a case where the pedestrians are highly overlapped. Each image pair in Fig.2(b) demonstrates a case where the same pedestrian in different images was annotated with different identity labels. Person search aims to obtain discriminate feature representation between different identities. Congestion and occlusion lead to an ambiguous overlapped area between different annotation boxes. The mislabeled pedestrians in the training set may encourage the network to discriminate pedestrians samples belonging to the same identity. Above mentioned cases encourage the network to pay more attention to illumination, viewpoints, and backgrounds in these image samples. This may lead to inter-identity similarity and intra-identity variation. As a result, the performance of person search models may be compromised.

For *scene-inherent* noise, we propose a context contrastive loss to extract noise-robust features. The philosophy of contrastive learning is utilized to compare pedestrians in scene images to ensure feature discrimination between different identities while maintaining the consistency between features from the same pedestrian identity. The network can thus learn to distinguish pedestrians under congestion and occlusion scenarios. A unique in-batch label can be further assigned to the unlabeled pedestrians and the unlabeled pedestrians can be included in the comparison. As a result, our methods can take full advantage of contextual information in scene images and assure the discrimination between not only labeled pedestrians but also unlabeled pedestrians. The difference between person re-identification, previous person search, and our methods are discussed in Fig.1. The training process of person re-identification only assures feature discrimination between different pedestrian identities. Previous person search approaches take a step forward and guarantee the discrimination between labeled identity and unlabeled pedestrians. In contrast, our method takes full advantage of contextual information in scene images and further constrains the feature similarity between unlabeled identities. While some methods also include contextual information, we introduce contextual information as supervision instead of a test-time assistance [20]–[22]. To the best of our knowledge, we are the first to introduce contextual information into the training process of person search. To learn more robust features, we further introduce the data uncertainty learning [23] into person search. The data uncertainty learning models the embedding as a probability distribution instead of a deterministic one to capture the inherent noise in the data and mitigate the negative influence of the noise so that more robust features can be obtained.

For *human-introduced* noise, we propose a noise-discovery and noise-suppression training schema, and a noise-suppression Online Instance Matching(OIM) [7] loss is proposed to suppress the noise from mislabeled pedestrians. The training schema includes a two-pass training process. The first training pass trains a vanilla person search network. After the first training pass, preliminary pedestrian features are obtained from the model trained in the first pass. Based on the preliminary pedestrian features, a feature prototype for each pedestrian identity is obtained from the pedestrian features of the same identity, and label noise can be discovered by analyzing the relation of feature prototypes from the different pedestrians. During the second training pass, the OIM loss is substituted with the noise-suppression OIM loss. The noise discovered will be suppressed by cutting off the gradient flow of mislabeled pedestrians during back-propagation. As a result, the unexpected gradient from the label noise will be stopped from interfering with the training process.

We summarize the *motivation* of this work as:

- The way of utilizing contextual information in the existing methods has some limitations and may not generalize to more realistic scenarios.
- Relationship between pedestrians in the same scene can be better utilized in the training process of person search for more discriminative features.
- Noise from congestion, occlusion, and annotation process may have a negative influence on the training process and compromise the performance of person search network.

The *contributions* of this work are three folds:

- A context contrastive loss is designed to learn more robust features under congestion and occlusion. To the best of our knowledge, we are the first to leverage contextual information from scene images as supervision during the training process.

- The data uncertainty learning is introduced to the person search to model and alleviate inherent noise in the training data for more robust feature learning.
- A noise discovery and noise suppression mechanism is proposed to alleviate the negative influence of label noise.

II. Related Work

A. Person Search

Person search bridges the gap between person re-identification and real-world application by integrating pedestrian detection and person re-identification into a unified framework. [6] was the first work to discuss the gap and raised interest in the community. Existing works of person search can be categorized into two technical approaches: *end-to-end methods* and *two-step methods*.

1) End-to-End Methods: The first end-to-end person search method was proposed in [7]. The hierarchical relationship between detection and re-identification was analyzed in [10]. [9] proposed to detangle the mutual feature of two subtasks into the norm and normalized feature for classification and identification. To mitigate the negative influence of feature misalignment, [13] proposed to align the partial features from detection with regression. [14] combined instance segmentation and keypoint detection into person search. [24] fused bounding box features from multiple backbone stages to obtain more discriminative features. [20] proposed to aggregate surroundings of pedestrians to include information like belongings and co-travelers. [25] proposed an online pairing loss with hard example mining for person search.

2) Two-Step Methods: Detection scores were integrated into pedestrian similarity measurement in [8]. [19] proposed to utilize mask as guidance information. The affine transformation was used as a differentiable RoI operation in [16] and the detector can thus be supervised with re-identification loss. A deeper analysis on two subtasks of person search was carried out in [22] and a similarity measurement base on objectness and surrounding pedestrians was proposed.

Also, some ideas have been explored with both approaches. Knowledge distillation was utilized to achieve better performance in [11], [15], [26]. Some other methods [12], [17], [18] adopted query identities as guidance during feature extraction. Some work also proposes to redefine person search as a detection-free process using reinforcement learning [27] or LSTM [28] to iteratively shrink the regions with target persons. However, these methods are complex but have only limited accuracy.

B. Person Re-Identification

Person re-identification aims at matching given pedestrians from massive pedestrian images. Early person re-identification methods are limited by the representability of hand-crafted features. With the renaissance of CNN, deep-learning-based methods have dominated the field of person re-identification. [29] proposed a refined part pooling to re-assign outliers in part models. [30] split pedestrian images into patches to learn features from multiple granularities. [31] proposed to use batch normalization neck to separate the conflict of metric loss and classification losses. [32] proposed to dynamically align and match local information. [33] fused triplet loss together with center loss and proposed a center-triplet loss. [34] proposed a comparative similarity loss based on pedestrian triplets and designed a multi-scale network to distinguish the pedestrians better. Some other person re-identification methods utilized extra information from video sequence [35], [36]. [37] focused on the misalignment of pedestrians and a pedestrian alignment network to learn more consistent pedestrians features. Facing a similar challenge, our work proposes to compare the pedestrian features in the same scene image. Some methods focus on novel topics like weak supervision [38], [39] and attack detection [40]. Meanwhile, some methods using GANs to bridge the domain gap between different cameras [41], [42] or generate more training data [43].

C. Usage of Contextual Information

Contextual information is utilized in a wide range of research areas. [44] proposed to use feature around the target to reweighs the class probability predicted in the task of object detection; [45] utilized the relation between different targets with attention; [46] designed a context loss for image transformation with semantic awareness; [47] take advantage of the contextual similarity between semantic and visual features to deal with sparse objects association across frames; [48] proposed a contextual bilateral loss that is robust to mild misalignment between input and outputs images.

D. Noise and Occlusion

Reducing the interference of noise and occlusion is a long-lasting research topic due to the actual demand in real-world applications. [49] proposed to use Expectation-Maximization algorithm to predict the noise level and true label with a small amount of clean data and massive data with noise. [50] proposed a noise-tolerant detector to mine the label and train the object detection model in a semi-supervised fashion. [51] utilized two CNNs to train each network with high confidence samples predicted by the other network. [52] employed compositional nets to model context and object representation separately and increase the robustness under occlusion. [53] add a bounding box occlusion estimator to the backbone and train it in an adversarial manner. While our work focuses on a similar challenge, the intention is different. [54] estimated the credibility of the pseudo-label and minimize the negative influence of noisy labels. Our work aims to learn discriminative person re-identification features that are robust to congestion, occlusion, and ambiguous samples.

III. Method

In this section, we will introduce CANR in detail. As a common approach for end-to-end person search, we adopt Faster R-CNN as our base detector. The overall framework of CANR is shown in Fig.3(a). CANR takes scene images as input and outputs the bounding boxes for pedestrian localization and the corresponding features for person re-identification.

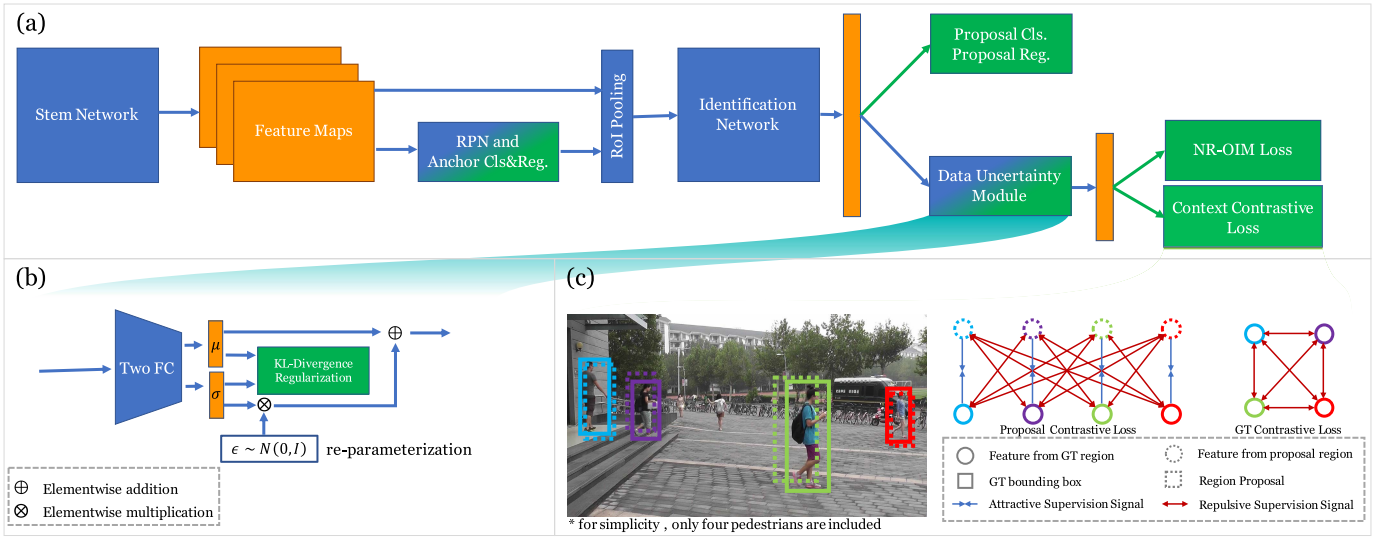


Fig. 3. Overall framework and details of key components of CANR. (a) CANR mainly consists of an end-to-end person search network, the data uncertainty module, and the context contrastive loss. (b) The data uncertainty module predicts the mean and variance of the feature embedding and models the feature embedding as a probability distribution. (c) The context contrastive loss compares each pedestrian with the other pedestrians to ensure the feature discrimination among different identities while maintaining feature consistency of the proposal regions assigned to the same ground truth bounding boxes.

The context contrastive loss and the data uncertainty learning will be introduced in section III-A and III-B respectively. The noise discovery and noise-suppression OIM will be introduced in section III-C.

A. Context Contrastive Loss

Inspired by contrastive learning, we propose *context contrastive loss* to learn congestion-robust and occlusion-robust features while fully utilize the information from scene images. It contains two sub-losses: *proposal context contrastive loss* and *ground truth context contrastive loss*.

Proposal context contrastive loss is illustrated in Fig.3(c). The proposal context contrastive loss compares each feature from proposal regions with the features from ground truth bounding boxes. It assures the consistency between feature from each proposal region and the ground truth region assigned to it. Meanwhile, it also guarantees the feature discrimination between features from each proposal region and the other ground truth regions. If the same identity exists in the multiple images within a batch, the ground truth features of the same identity will be summed and normalized to unit length.

The proposal context contrastive loss for feature $\boldsymbol{x}$ from a given positive proposal region can be acquired by:

$$L_{p\_cc} = -\log\left(\frac{\exp\left(\boldsymbol{g}_i^{\top}\boldsymbol{x}\cdot s_p\right)}{\sum_{j=1}^{G}\exp\left(\boldsymbol{g}_j^{\top}\boldsymbol{x}\cdot s_p\right)}\right) \tag{1}$$

where $\boldsymbol{x}$ represents the feature of a proposal region assigned to the i-th ground truth bounding box, $\boldsymbol{g}_i$ and $\boldsymbol{g}_j$ denote feature of the i-th and j-th ground truth bounding boxes respectively, G denotes the number of annotated pedestrian boxes in the image, and s_p is the scale factor for proposal context contrastive loss. Both $\boldsymbol{g}_i$ and $\boldsymbol{x}$ are unit length normalized. Intuitively, the loss can be seen as a L_2-constrained bias-free softmax cross-entropy loss where the features of ground truth bounding boxes act as the weight, and the features of proposal regions are classified into different pedestrians.

To avoid the noisy features of proposal regions from interfering with features of ground truth regions. We design a gradient scaler $f\left(x,\ s_{scaler}\right)$ to scale the gradient from the ground truth features by a scale s_{scaler}, the property of the gradient scaler can be expressed as:

$$f\left(\boldsymbol{x},\ s_{scaler}\right) = \boldsymbol{x} \tag{2}$$

$$\frac{\mathrm{d}f\left(\boldsymbol{x}, s_{scaler}\right)}{\mathrm{d}\boldsymbol{x}} = s_{scaler}\cdot \mathrm{d}f\left(\boldsymbol{x}, s_{scaler}\right) \tag{3}$$

where $\boldsymbol{x}$ is input. During the forward propagation, the scaler just outputs the original input feature. During back-propagation, the gradient is scaled by s_{scaler}.

To increase the number of proposals and facilitate downstream proposal context contrastive loss, the ground truth bounding boxes are jittered for more proposal regions. The coordinates of the jittered bounding boxes are obtained by:

$$x_{j_i} = x_{gt_i} + N\left(0, \left(\lambda_s\left(x_{gt_2} - x_{gt_1}\right)\right)^2\right) \tag{4}$$

$$y_{j_i} = y_{gt_i} + N\left(0, \left(\lambda_s\left(y_{gt_2} - y_{gt_1}\right)\right)^2\right) \tag{5}$$

where $N\left(\mu, \sigma^2\right)$ denote the Normal distribution with μ as mean and σ as standard deviation; $\left(x_{gt_1}, y_{gt_1}\right)$ and $\left(x_{gt_2}, y_{gt_2}\right)$ denote the top left and the bottom right coordinates of the ground truth bounding boxes; $\left(x_{j_1}, y_{j_1}\right)$ and $\left(x_{j_2}, y_{j_2}\right)$ denote the top left and bottom right coordinate of the jittered ones, λ_s denotes the magnitude of the disturbance.

The *ground truth context contrastive loss* is illustrated in Fig.3(c). It constrains the discrimination between features from different ground-truth regions. Similar to the proposal context contrastive loss, each feature from ground truth pedestrian bounding boxes is compared against features from the other

ground truth bounding boxes. The ground truth contrastive loss for ground-truth feature $\boldsymbol{g}_i$ is acquired by:

$$L_{gt\_cc} = -\log\left(\frac{\exp\left(\boldsymbol{g}_i^\top \boldsymbol{g}_i \cdot s_g\right)}{\sum_{j=1}^{G}\exp\left(\boldsymbol{g}_j^\top \boldsymbol{g}_i \cdot s_g\right)}\right) \tag{6}$$

where s_g is the scale for ground truth context contrastive loss. L_{gt_cc} is designed to supervise the network to be more discriminative to the neighboring pedestrians and similar environment.

Moreover, to learn crowd-robust pedestrian features and facilitate the discrimination between neighboring pedestrians, we insert a spatial attention module into the identification network. The context contrastive loss can act as stronger supervision over spatial attention compared with conventional use cases. The spatial attention module can then learn to emphasize the informative regions with pedestrians and suppress the distraction from neighboring pedestrians and the background. The spatial attention module can be expressed as:

$$\mathbf{F}_{att} = f_\sigma\left(f_{7\times7}([\text{AvgPool}(\mathbf{F}); \text{MaxPool}(\mathbf{F})])\right) \odot \mathbf{F} \tag{7}$$

where $\mathbf{F}_{att}$, f_σ, $f_{7\times7}$, $\odot$ and $\mathbf{F}$ denote the feature map after spatial attention, sigmoid activation function, convolution with kernel size of 7×7, elementwise multiplication and the input feature map.

While L_{p_cc} and L_{gt_cc} have similar formulation, they are designed based on different intuition. L_{p_cc} is designed to assure the discrimination of pedestrian features even though the proposal regions do not cover the target pedestrians precisely and include distraction from background or the other pedestrians. L_{gt_cc} is designed to constrain feature discrimination under congestion situations. With the context contrastive loss, both labeled and unlabeled pedestrians can be utilized to train the network to extract congestion robust features. Moreover, with the context contrastive loss, the model can also learn to distinguish the pedestrians in the same image more directly and intuitively. With an in-batch label assigned to the unlabeled pedestrians, the context contrastive loss can be applied to the batch level and the context information and the unlabeled person can be better utilized during training.

B. Data Uncertainty Module

To further minimize the negative influence of inherent noise in the data and learn noise-robust features, we introduce the data uncertainty learning into person search. The data uncertainty learning captures the inherent noise in the data by modeling the feature and the uncertainty simultaneously. Specifically, a given feature from the identification network is modeled as a Gaussian distribution, w.r.t the $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. The $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are predicted by two separate feature projectors by:

$$\boldsymbol{\mu}_i = f_{\boldsymbol{\theta}_\mu}(\boldsymbol{z}_i) \tag{8}$$

$$\boldsymbol{\sigma}_i = f_{\boldsymbol{\theta}_\sigma}(\boldsymbol{z}_i) \tag{9}$$

where $\boldsymbol{z}_i$ is feature embedding of the i-th sample from the identification network, $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ are the mean and variance predicted by the feature predictors respectively. The $\boldsymbol{\theta}_\mu$ and $\boldsymbol{\theta}_\sigma$ are the parameters for the feature projectors.

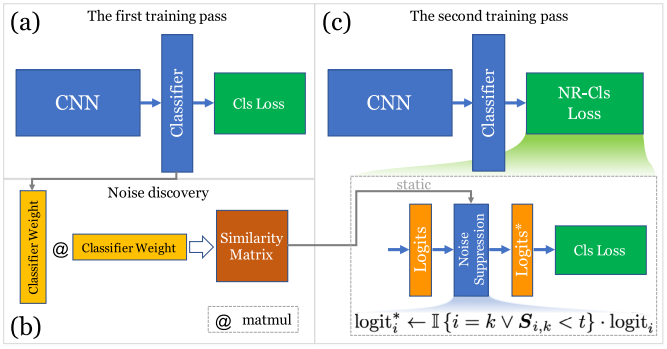


Fig. 4. Illustration of the first training pass (a), noise discovery (b), and the second training pass (c). Label noise is discovered after the first training pass and suppressed in the second training pass.

As a result, the embedding of a given pedestrian is a distribution instead of a deterministic embedding. During training, the feature embedding for the loss functions is sampled from the predicted distribution using the re-parameterization [55] to ensure the differentiability for back-propagation. During inferencing, $\boldsymbol{\sigma}_i$ is omitted and only the $\boldsymbol{\mu}_i$ is adopted for similarity measurement. The final features for person re-identification are acquired by:

$$\boldsymbol{x}_i = \begin{cases} \boldsymbol{\mu}_i + \boldsymbol{\epsilon}\boldsymbol{\sigma}_i & \text{if training} \\ \boldsymbol{\mu}_i & \text{otherwise} \end{cases} \tag{10}$$

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{I})$ and $\boldsymbol{x}_i$ denotes the i-th feature for person re-identification.

To prevent the prediction of $\boldsymbol{\sigma}_i$ from collapsing into zero, making distribution-based embedding degrade into the deterministic embedding, KL divergence is utilized to constrain the $N\left(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2\right)$ into normal distribution:

$$\begin{aligned} L_{KL} &= KL\left[N\left(\boldsymbol{x}_i \mid \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2\right) \middle\| N(\boldsymbol{\epsilon} \mid \mathbf{0}, \boldsymbol{I})\right] \\ &= \frac{1}{2}\sum_{j=1}^{D}\left(\sigma_{i_j}^2 + \mu_{i_j}^2 - 1 - \ln\left(\sigma_{i_j}^2\right)\right) \end{aligned} \tag{11}$$

where σ_{i_j}, μ_{i_j}, D denote the j-th dimension of $\boldsymbol{\sigma}_i$, the j-th dimension of $\boldsymbol{\mu}_i$ and the dimension of feature vector.

While the idea of data uncertainty learning has been explored in the task of person re-identification [56], the formulation of the data uncertainty in [56] is not suitable for the task of person search. The main difference between person re-identification and person search is that the ROI region may contain background areas. However, the uncertainty formulation in [56] does not explicitly constrain these background samples to have a suitable uncertainty. This may lead to large uncertainty for the background regions and interfere with the uncertainty learning of foreground regions. By contrast, the L_{KL} in our method explicitly constrains the prediction of uncertainty under a relational range, leading to more stable uncertainty learning.

C. Noise Discovery and Noise Suppression

To minimize the negative influence of mislabeled pedestrians during the training phase, a noise-aware training process is

proposed to discover and suppress the mislabeled pedestrians. It includes two separate training passes. The first pass discovers the mislabeled pedestrians and the second pass cuts off the gradient flow of the mislabeled pedestrians. A simplified process is illustrated in Fig.4.

As illustrated in Fig.4(a), the first pass trains a person search network with vanilla OIM loss. After the first training pass, the potential mislabeled pedestrians in the training set is analyzed and discovered as shown in Fig.4(b). The feature prototype of each pedestrian is obtained for similarity comparison. Thanks for the momentum update of the lookup table in the OIM loss, each feature vector in the lookup table can be seen as the feature prototype of each pedestrian. As the result, we simply take the weight of the lookup table in OIM loss $\boldsymbol{V} \in \mathbb{R}^{L\times D}$ as the feature prototype, where L denotes the number of pedestrian identities and D denotes the dimension of the feature embedding. The similarity matrix $\boldsymbol{S} \in \mathbb{R}^{L\times L}$ between different feature prototypes in the training set can be acquired by $\boldsymbol{S} = \boldsymbol{V}\boldsymbol{V}^{\top}$. The similarity matrix $\boldsymbol{S}$ measures the similarities between different pedestrian identities in the training set. Fig.4(c) demonstrates the noise suppression during the second training pass, the logits predicted by the network are dynamically cut off based on the similarity matrix $\boldsymbol{S}$ in the classification loss (OIM in our case) by, (12), shown at the bottom of the page, where p_i denotes the probability of $\boldsymbol{x}$ being classified into the i-th identity, $\mathbb{I}\{\cdot\}$ denotes the indicator function, k denotes the target label of $\boldsymbol{x}$, $\boldsymbol{S}_{i,k}$ denotes the similarity between the i-th identity and the k-th identity, τ is the temperature for the OIM loss, $\boldsymbol{v}_j$ denotes the j-th feature in the LUT, $\boldsymbol{u}_m$ denotes the m-th feature in the CQ, L denotes the size of LUT, Q denotes the size of CQ, and t is a hyper-parameter for threshold. We recommend referring to [7] for a detailed explanation of LUT and CQ.

Next we elaborate the role of the noise supression from the view of derivation. For ease of illustration, we discuss a simplified classification case with softmax cross entropy loss where the bias are omitted. The score of class c is obtained by $s_c = \mathbb{I}\{\cdots\}\boldsymbol{w}_c\boldsymbol{x}$ where $\boldsymbol{x}$ denotes the input, $\boldsymbol{w}_c$ denotes classifier weight of class c and $\mathbb{I}\{\cdots\}$ denotes the indicator mentioned in Eqn. 12. The derivation w.r.t. $\boldsymbol{w}_c$ and $\boldsymbol{x}$ can be obtained by:

$$\frac{\partial L_{cls}}{\partial \boldsymbol{w}_c} = \begin{cases} \dfrac{\exp(s_c)}{\sum_k \exp(s_k)}\boldsymbol{x}\mathbb{I}\{\cdots\} & y \neq c \\ -\boldsymbol{x} + \dfrac{\exp(s_c)}{\sum_k \exp(s_k)}\boldsymbol{x} & y = c \end{cases} \tag{13}$$

$$\frac{\partial L_{cls}}{\partial \boldsymbol{x}} = -\boldsymbol{w}_y + \sum_c \frac{\exp(s_c)\boldsymbol{w}_c\mathbb{I}\{\cdots\}}{\sum_l \exp(s_l)} \tag{14}$$

where L_{cls} is obtain by $L_{cls} = -\ln(\frac{\exp(s_y)}{\sum_k \exp(s_k)})$, and y is the ground truth label for the label $\boldsymbol{x}$.

To illustrate the above equations with an example, we assume there is a sample $\boldsymbol{x}$ of a specific identity with the label y_t. The same identity is also mislabeled as y_n in some images. The $\frac{\exp(s_{y_t})}{\sum_k \exp(s_k)}$ in the $\frac{\partial L_{cls}}{\partial \boldsymbol{w}_{y_t}}$ will not be interferenced by the score of noise label s_{y_n}. As a result, $\frac{\exp(s_{y_t})}{\sum_k \exp(s_k)}$ will be much closer to the value without noise. Thus, the update of $\boldsymbol{w}_{y_t}$ would be more stable during the training stage. The $\frac{\partial L_{cls}}{\partial \boldsymbol{w}_{y_n}}$ will be 0, isolating the interference between $\boldsymbol{x}$ and $\boldsymbol{w}_{y_n}$. Also, $\frac{\partial L_{cls}}{\partial \boldsymbol{x}}$ will not include $\frac{\exp(s_{y_n})\boldsymbol{w}_{y_n}}{\sum_l \exp(s_l)}$. In this way, the network can prevent the undesired gradient of noisy labels from interfering with the other parameters and, as a result, minimize the negative influence of mislabeled pedestrians.

D. Training and Loss Formulation

CANR includes a two-pass training process. Generally, the purpose of the first training pass is to discover the noise and the second training pass suppress the noise discovered in the first training pass to achieve better accuracy. The first training process trains a model without noise suppression. After the first training process, noise in the training set is analyzed by comparing the feature prototype between different identities. During the second training pass, a model with noise suppression is trained. The noise suppression dynamically cut off the gradient flow of the mislabeled samples based on the noise discovered in the first training pass. Besides the noise suppression introduced in the second training process, the two training passes are identical. The training process of CANR is demonstrated in the Algorithm 1.

The total loss for our end-to-end person search network is defined as:

$$L = L_{det} + L_{OIM} + L_{p\_cc} + L_{gt\_cc} + \lambda_{KL} \cdot L_{KL} \tag{15}$$

where the L_{det} represents the losses from Faster R-CNN, λ_{KL} is the weighting factor for L_{KL} and the L_{OIM} represents the online instance matching loss from [7].

E. Miscellaneous

1) Adaptive Pooling Fusion: Max pooling and average pooling both have drawbacks [57]. Adaptive Pooling Fusion is proposed to fuse the features from max pooling and average pooling. The final feature representation is acquired by:

$$\boldsymbol{z} = \lambda_w \boldsymbol{z}_{avg} + (1 - \lambda_w)\, \boldsymbol{z}_{max} \tag{16}$$

where $\boldsymbol{z}_{avg}$ and $\boldsymbol{z}_{max}$ denote the features acquired by average pooling and max pooling from the identification network respectively. λ_w is a weighting factor to balance features from different pooling strategies, which is acquired by $\lambda_w = \text{sigmoid}\,(p)$. p is a learnable parameter and is updated through back-propagation.

$$p_i = \frac{\exp\left(\mathbb{I}\left\{i = k \vee \boldsymbol{S}_{i,k} < t\right\}\boldsymbol{v}_i^{\top}\boldsymbol{x}/\tau\right)}{\sum_{j=1}^{L}\exp\left(\mathbb{I}\left\{j = k \vee \boldsymbol{S}_{j,k} < t\right\}\boldsymbol{v}_j^{\top}\boldsymbol{x}/\tau\right) + \sum_{m=1}^{Q}\exp\left(\boldsymbol{u}_m^{\top}\boldsymbol{x}/\tau\right)} \tag{12}$$

Algorithm 1 Training Phase of CANR

Input:
Training set S, total epoches $epoches$
Output:
Trained Model weight $\boldsymbol{W}$

Initialize the model weight $\boldsymbol{W}$
for $pass$ in {1, 2} **do**
  for e in {1, 2, ..., $epochs$} **do**
    for batch B sampled from S **do**
      Extract pedestrian features Z_{ped} from B
      Calculate detection related losses, get L_{det}
      Feed Z_{ped} into the Data Uncertainty Module, get F_{ped} and L_{KL} (Eqn. 11)
      Construct contrastive pairs and calculate the context contrastive loss, get L_{p_cc} (Eqn. 1) and L_{gt_cc} (Eqn. 6)
      if $pass = 1$ **then**
        Feed F_{ped} into the Online Instance Matching (OIM) loss, get L_{OIM}
      end if
      if $pass = 2$ **then**
        Suppress the noise from F_{ped} with P_{noise}, calculate the OIM loss, get L_{OIM}
      end if
      Calculate total loss L (Eqn. 15), back propagation, and update $\boldsymbol{W}$
    end for
  end for
  if $pass = 1$ **then**
    Analysis the noise of the dataset and set a set of noise pairs of identities P_{noise}
  end if
end for
return trained model weight $\boldsymbol{W}$

2) Deformable RoI Pooling: Proposal regions from the RPN in the faster R-CNN tend not to include all body parts of the target pedestrian. It's suboptimal for the fine-grained task like person re-identification. To alleviate this issue, we introduce the Deformable RoI Pooling into our work. Deformable RoI Pooling learns a sample point offset for each position on the output feature map. The context contrastive loss can also act as guidance over the training of the offset prediction. In this way, our person search network will learn to focus on the target pedestrians and will be more robust to low-quality region proposals.

F. Discussion

1) Discussion on the Triplet Loss and the Context Contrastive Loss: The triplet loss enforces a margin between the similarity of positive pairs and negative pairs. The triplet loss requires to sample p identities with k images for each identity to form a batch during training. The sampling process ignores the contextual information that comes with scene images. In contrast, our method takes advantage of the contextual information, neighboring pedestrians under congested scenarios can be compared directly, and features robust to congestion and occlusion can be learned. While we adopt softmax cross-entropy to constrain the feature discrimination of pedestrians in the same scene, triplet-style loss functions can also be applied for a similar purpose. Furthermore, the context contrastive loss can be used in parallel with most loss functions.

2) Discussion on Hard Examples and the Label Noise: The hard examples are supposed to be discriminated in the first training phase of the noise discovery and noise suppression process. Also, there is no point to include the extreme hard example pairs if the ordinary training process can't distinguish them well. We find that the vast majority of mislabeled pedestrians have two labels. As a result, only the maximal logit of a given example satisfy $\mathbb{I}\{i \neq k \wedge \boldsymbol{S}_{i,k} > t\}$ is zeroed out in practice. We empirically find that 0.75 is a suitable cosine distance threshold to distinguish label noise and severe overlapped pedestrians (illustrated in Fig.2(a)) from hard examples on both datasets. For severe overlapped pedestrians, they are always in the same scene images and the context contrastive loss will handle this situation.

3) Discussion on the Effectiveness of Data Uncertainty Learning: To simplify the discussion, we treat the person re-identification sub-task of person search as an isolated task which includes L_{ReID} (w.r.t. L_{OIM}, L_{p_cc} and L_{gt_cc} in our work) and L_{KL}. The objective of L_{ReID} is to supervise the network to predict discriminative person re-identification features for each sample. The objective of L_{KL} to constrain the $\boldsymbol{\mu}_i$ to have zero meanwhile restricting the $\boldsymbol{\sigma}_i$ neither collapsing to 0 nor too large. Constraining the $\boldsymbol{\mu}_i$ to have zero mean can help the features to distribute more evenly on the hypersphere. For the clean samples, predicting relatively small $\boldsymbol{\sigma}_i$ may facilitate lower L_{ReID} and lead to optimal overall loss. Noise samples would result in large L_{ReID} and giving such samples larger $\boldsymbol{\sigma}_i$ may facilitate better overall loss. Moreover features sampled from a distribution with large $\boldsymbol{\sigma}_i$ may cancel each other out and, as a result, have less impact on the training compared with the ones with small $\boldsymbol{\sigma}_i$. From a different perspective, the data uncertainty learning can also be seen as a kind of regularization where a feature after disturbance should still be discriminative enough to recognize the identity.

4) Comparison With Our Previous Work: Our previous work [58] proposed a loss function that aims to achieve better feature discrimination between different pedestrians and improve feature consistency of the same identity. An attention module is also proposed to help the network suppress background clutters and stress the pedestrians. This work focus on the scene-inherent noise and human-introduced noise. Specifically, a context contrastive loss is proposed to suppress interference from surrounding pedestrians under congested scenarios; data uncertainty learning is introduced into person search for better noise-robust person search. Also, noise discovery methods and noise suppression loss are proposed to mitigate the negative influence of mislabeled pedestrians.

IV. Experiments

A. Dataset and Evaluation Protocol

a) The CUHK-SYSU dataset: Is collected from street snaps shoot by handheld cameras and movie snapshots. It features diverse scene variations and large quantities of pedestrian identities. It has a total of 18,184 images, 43,110 bounding box annotations, and 8,432 identity labels. The training set of the dataset has 11,206 images and 5,532 identity labels while the testing set has 6,978 gallery images and 2,900 query pedestrians.

b) The PRW dataset: Contains images from video streams of six fixed cameras on campus. It features different camera styles and great scale variation. It has a total of 11,816 images, 34,304 bounding box annotations, and 932 identity labels. The training set has 5,704 images and 482 identities while the testing set has 6,112 gallery images and 2,057 query pedestrians.

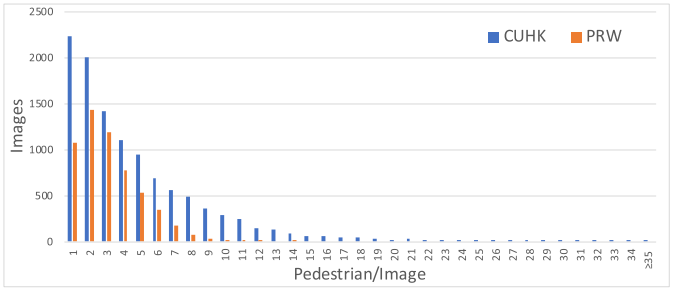
c) Evaluation metrics: We adopt the cumulative matching characteristics (CMC) and mAP as the evaluation metrics, which are commonly used in person search. Both these two metrics are inherited from the task of re-identification. A match is considered positive when IoU between the predicted bounding box and the ground truth one is larger than 0.5.

B. Implementation Details

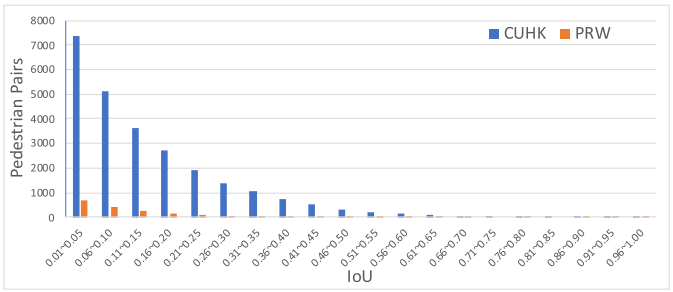
We implemented CANR based on Faster R-CNN on top of MMDetection with PyTorch. The ResNet-50 was adopted as the backbone. The first four blocks (*layer0* to *layer3*) of ResNet-50 were adopted as the stem network and the last block (*layer4*) was adopted as the re-identification network. On top of the feature from the stem network, the Regional Proposal Network [62] was utilized to extract potential regions containing pedestrians. The extracted proposal regions were sampled and then pooled using RoI Align [63]. The feature maps obtained from the ROI align were then sent to the identification network. The Online Instance Matching (OIM) loss, the context contrastive loss, and the KL divergence loss were utilized to supervise the re-identification sub-task. Meanwhile, losses of Faster R-CNN were utilized to supervise the detection sub-task. The model was trained with two NVIDIA RTX 2080Ti in distributed data-parallel or an RTX 3090 with equivalent settings. We follow the standard 2x schedule in object detection, the batch size was set to 4 for each GPU, the learning rate is set to 4e-3, and the weight decay was set to 1e-4. The λ_{KL} is set to 1e-3 according to [23]. For the PRW dataset, the s_p was set to 8 and s_g was set to 8. Given that the number of pedestrians varies dramatically in different images of the CUHK-SYSU dataset, we refer to [64] and empirically set s_p to $5\sqrt{2}\ln\left(n_{gt}\right)$, where n_{gt} is the number of ground truth regions in each batch. The ground truth contrastive loss and bounding box jittering are empirically omitted for the CUHK-SYSU dataset. For the rest of the parameters, we followed the settings of NAE [9] and MMDetection.

C. Comparison With State-Of-the-Art Methods

To evaluate the performance of CANR, the performance of CANR is compared against the state-of-the-art methods on the CUHK-SYSU and the PRW datasets. Existing state-of-the-art methods are categorized into two-step methods [15], [16], [18], [19], [22] and end-to-end methods [7], [9]–[14], [20], [21], [24], [28], [59]–[61]. The results involved in the comparison are taken from the original papers. Some of the state-of-the-art methods rely on external information like key points, masks, and knowledge distillation from an external network. To make the comparison fair, we present the extra information required for the state-of-the-art methods. 'Keypoints' and 'Mask' means key points or masks are required during training; 'Knowledge Distillation' means feature-level supervision from an external network is utilized, 'Detection Confidence' means the detection confidence weighted similarity is used during the retrieval process. It should also be noted that DMRN [59] and AlignPS [61] are based on stronger detectors. AlignPS [61] also combines features from different scales for a more robust pedestrian feature representation.



(a)



(b)

Fig. 5. Statistics of (a) pedestrians per image in both dataset, (b) IoU of pedestrian pairs.

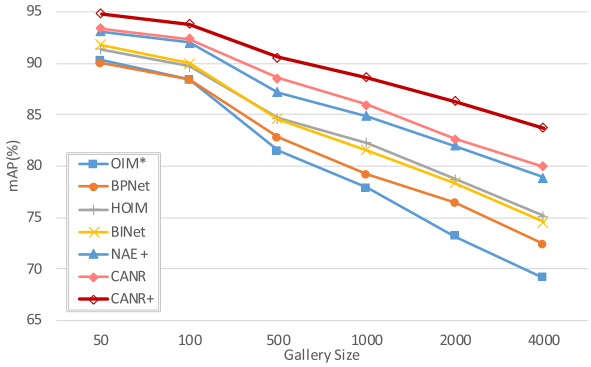


Fig. 6. The mAP of CANR and CANR+ on the CUHK-SYSU under different gallery sizes. Compared to existing end-to-end person search works, CANR+ achieves the best performance under different gallery sizes.

TABLE I
PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE CUHK-SYSU AND PRW DATASETS

Methods	Publication	Additional Information Required for ReID	CUHK-SYSU		PRW	
			mAP	Top-1	mAP	Top-1
Two-Step Methods						
MGTS[19]	ECCV 18	Mask	83.00	83.70	32.60	72.10
CLSA[15]	ECCV18	None	87.20	88.50	38.70	65.00
RDLR[16]	ICCV19	None	93.00	94.20	42.90	70.20
TCTS[18]	CVPR20	None	**93.90**	**95.10**	**46.80**	**87.50**
Faster R-CNN + PCB + OR[22]	TIP21	None	92.93	93.69	43.01	65.87
End-to-End Methods						
OIM[7]	CVPR17	None	75.50	78.70	21.30	49.90
NPSM[28]	ICCV17	None	77.90	81.20	24.20	53.10
QEEPS[12]	CVPR19	None	88.90	89.10	37.10	76.70
GRAPH[21]	CVPR19	None	84.10	86.50	33.40	73.60
BPNet[14]	AAAI20	Keypoint&Mask	88.40	90.50	48.50	**87.90**
HOIM[10]	AAAI20	None	89.74	90.83	39.77	80.36
APNet[13]	CVPR20	Keypoint	88.90	89.30	41.90	81.40
BINet[11]	CVPR20	Knowledge Distillation	90.30	91.40	47.20	87.00
NAE[9]	CVPR20	Detection Confidence	91.50	92.40	43.30	80.90
NAE+[9]	CVPR20	Detection Confidence	92.10	92.90	44.00	81.10
BUFF[24]	MM20	None	90.70	91.60	42.20	81.00
DCAR[20]	MM20	None	87.50	88.70	38.80	77.70
IIDFC[58]	TCSVT21	None	91.96	93.34	43.43	83.37
DMRN[59]	AAAI21	None	93.20	94.20	46.90	83.30
DKD[60]	AAAI21	Knowledge Distillation	93.09	94.24	**50.51**	87.07
AlignPS[61]	CVPR21	None	93.10	93.40	45.90	81.90
AlignPS+[61]	CVPR21	None	**94.00**	94.50	46.10	82.10
CANR(Ours)	TCSVT22	None	92.39	93.21	43.41	83.81
CANR+(Ours)	TCSVT22	None	93.86	**94.52**	44.78	83.86

The results are presented in Table.I. For the CUHK-SYSU dataset, CANR achieves 92.39% on mAP and 93.21% on Top-1. With deformable ROI pooling, CNAR+ achieves 93.86% on mAP and 94.52% on Top-1. For the PRW dataset, CANR achieves 43.41% on mAP and 83.81% on Top-1, CANR+ achieves 44.78% on mAP and 83.86% on Top-1. Without the requirement of additional information, CANR and CANR+ achieve state-of-the-art-comparable performance on the CUHK-SYSU dataset and competitive performance on the PRW dataset. To investigate this phenomenon, we analyzed the statistical information of both datasets. Specifically, the pedestrian bounding box pairs and pedestrian per image is evaluated to indicate the extent of congestion in the dataset. The results are shown in Fig.5. Compared with the PRW dataset, the CUHK-SYSU dataset has a more congested scenario. Remind that one of our motivations is to focus on the scene inherited noise originated from congestion and occlusion. As a result, CANR and CANR+ will perform better in the congested scenario and this may account for the superior performance compared to the other methods on the CUHK-SYSU dataset. While the performance is not as competitive on the PRW dataset, CANR and CANR+ also bring satisfying improvement on top of the baseline and achieves competitive performance.

The performance under different gallery sizes is also evaluated on the CUHK-SYSU dataset. The performance is compared against state-of-the-art end-to-end person search methods and the results are shown in Fig.6. As the gallery size increases, the performance of all methods gradually decreases, which indicates that the task of person search is more challenging under larger gallery sizes. Our method achieves the best performance across different gallery sizes, which shows the effectiveness of CANR and CANR+.

TABLE II
ABLATION STUDY OF EACH COMPONENT OF CANR ON THE PRW DATASET. 'OIM*' DENOTES THE OIM RE-IMPLEMENTED, 'BSL' DENOTES THE BASELINE BUILT ON TOP OF OIM*, 'DUM' DENOTES THE DATA UNCERTAINTY MODULE, 'CCL' DENOTES THE CONTEXT CONTRASTIVE LOSS, 'PF' DENOTES THE POOLING FUSION, 'BJ' DENOTES THE BOUNDING BOX JITTERING, 'NS' DENOTES THE NOISE-SUPPRESSION LOSS, 'DP' DENOTES THE DEFORMABLE POOLING

OIM*	BSL	DUM	CCL	PF	BJ	NS	DP	mAP(%)	Top-1(%)
✓	✗	✗	✗	✗	✗	✗	✗	38.18	77.10
✓	✓	✗	✗	✗	✗	✗	✗	39.15	79.92
✓	✓	✓	✗	✗	✗	✗	✗	41.67	80.80
✓	✓	✓	✓	✗	✗	✗	✗	42.48	82.21
✓	✓	✓	✓	✓	✗	✗	✗	42.68	82.55
✓	✓	✓	✓	✓	✓	✗	✗	43.58	83.18
✓	✓	✓	✓	✓	✓	✓	✗	43.41	83.81
✓	✓	✓	✓	✓	✓	✓	✓	**44.78**	**83.86**
✓	✓	✗	✗	✗	✓	✗	✗	39.61	79.29

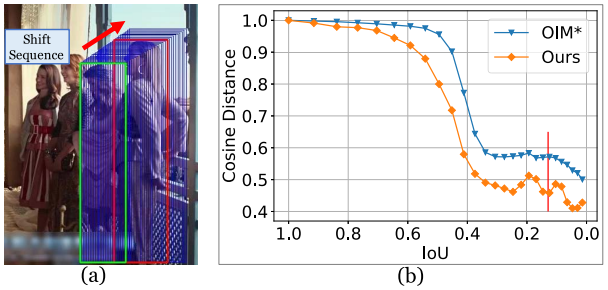
D. Ablation Study

To evaluate the effectiveness of each component, ablation studies on each component were conducted on the PRW and CUHK-SYSU datasets. The results are shown in Table.II and Table.III. We re-implement the OIM [7] and the OIM re-implemented is denoted as OIM* in the following part of this paper. Spatial attention and BNNeck are further added to act as the baseline for our method. For the PRW

TABLE III

ABLATION STUDY OF EACH COMPONENT OF CANR ON THE CUHK-SYSU DATASET

OIM*	BSL	PF	CCL	NS	DUM	DP	mAP(%)	Top-1(%)
✓	✗	✗	✗	✗	✗	✗	88.41	88.38
✓	✓	✗	✗	✗	✗	✗	90.90	91.62
✓	✓	✓	✗	✗	✗	✗	91.28	91.69
✓	✓	✓	✓	✗	✗	✗	91.79	92.48
✓	✓	✓	✓	✓	✗	✗	91.81	92.66
✓	✓	✓	✓	✓	✓	✗	92.39	93.21
✓	✓	✓	✓	✓	✓	✓	**93.86**	**94.52**



(a) (b)

Fig. 7. Comparison of shift sensitiveness between OIM and CANR. The green box and the red box denote the ground truth annotation of the targeted pedestrian and the neighboring pedestrian. The blue ones indicate the shifted boxes. Compared with the OIM, CANR tends to be more discriminative to inaccurate proposal regions. CANR is also more discriminative between neighboring pedestrians.

dataset, the OIM* achieves 38.18% and 77.10% in Top-1. The baseline gets 39.15% in mAP and 79.92% in Top-1. This setting is adopted as the baseline of our method. With the data uncertainty module, we get 41.67%(↑2.52%) in mAP and 80.80%(↑0.88%) in Top-1. On top of that, we add context contrastive loss and get 42.48%(↑0.81%) in mAP and 82.21%(↑1.41%) in Top-1. We further add the pooling fusion module and get 42.68%(↑0.20%) in mAP and 82.55%(↑0.34%) in Top-1. With bounding box jittering as data augmentation, we get 43.58%(↑0.90%) in mAP and 83.18%(↑0.63%) in Top-1. With the noise-suppression loss, we get almost identical mAP and 83.81%(↑0.63%) in Top-1. With all parts of CANR+, we get 44.78%(↑1.37%) in mAP and 83.86%(↑0.04%) in Top-1, which is a competitive result among end-to-end person search methods. The CUHK-SYSU dataset covers a wide variety of scenarios and many scene images contain abundant pedestrians. Therefore, there are enough positive samples for sampling. In this case, using BJ cannot bring positive effects and may disturb the data distribution of the proposal regions. As a result, we discarded the bounding box jittering on the CUHK-SYSU dataset. Based on the ablation study on the CUHK-SYSU dataset, similar results can be observed. To investigate the influence of the gradient scaler. We compare the scale gradient strategy with the stop gradient strategy and use the intact gradient strategy on the CUHK-SYSU dataset. Stop gradient prevents the gradient from back-propagating to previous parts while using intact gradient is equivalent to removing the gradient scaler. The results are shown in Table IV. Compared to the other settings, the scale gradient strategy achieves the best performance.

TABLE IV

RESULTS OF ANALYTICAL EXPERIMENT ON GRADIENT SCALER

Setting	mAP(%)	Top-1(%)
Stop Gradient	92.29	92.90
Gradient Scaler	92.39	93.21
Intact Gradient	92.31	92.97

TABLE V

COMPARISON OF DETECTION AND RE-IDENTIFICATION PERFORMANCE ON THE CUHK-SYSU DATASET

Method	Detection		ReID	
	Recall(%)	AP(%)	mAP(%)	Top-1(%)
Faster R-CNN	91.27	88.97	N/A	N/A
OIM*	90.34	87.44	88.41	88.38
CANR(Ours)	92.18	89.27	92.39	93.21
CANR+(Ours)	**92.59**	**89.77**	**93.86**	**94.52**

To evaluate the localization performance of CANR. Faster R-CNN, which is our base detection method, was trained on the CUHK-SYSU dataset. We keep the parameter setting from MMDetection, a standard 2x training schedule was adopted for training and the batch size was aligned with CANR. OIM [7] is re-implemented and evaluated on the CUHK-SYSU dataset. The hyper-parameter of OIM* was aligned with CANR for comparability. The results are shown in Table.V. CANR achieves 92.18% in recall and 89.27% in average precision(AP), which is a 0.91% and 0.30% performance improvement on recall and AP respectively compared with Faster R-CNN. This shows that CANR can facilitate the detection sub-task while achieving competitive performance on the re-identification sub-task. With CANR+, a slightly better detection can be further obtained. The OIM* achieved a slightly lower detection performance, which confirms the conflict [9] between two sub-tasks of person search. Compared with OIM*, CANR not only achieves better performance on detection but also outperforms OIM* on the re-identification task by a large margin, which shows the effectiveness of our method.

Further analytical experiments are conducted to understand the underlying reasonability of CANR. We shift the ground truth bounding box of a pedestrian and generate a group of boxes between two neighboring pedestrians. This process is illustrated in the left part of Fig.7(a). The generated boxes are selected as the proposal regions for the RoI Align. Then, features corresponding to these regions are obtained. Cosine distance is calculated between the features from ground truth bounding boxes and the generated boxes. The result is shown in Fig.7(b) and the red line indicates the IoU between two ground truth boxes. Compared with the OIM*, CANR is more discriminative to inaccurate proposal regions. CANR also gives a lower cosine distance between two adjacent pedestrians. Through this experiment, we may conclude that CANR is robust to inaccurate region proposals while more discriminative between neighboring pedestrians, which shows the effectiveness of CANR.

We also compare the runtime of CANR with different methods and the results are shown in Table.VI. The CANR

Fig. 8. Visualization of retrieval results. The blue boxes represent pedestrians to be queried, the green boxes show correct matches, and the red boxes indicate incorrect matches. Compared with NAE and OIM, CANR is more robust to extreme illumination condition (a), illumination changes (b), congestion (a), and occlusion (c, d).

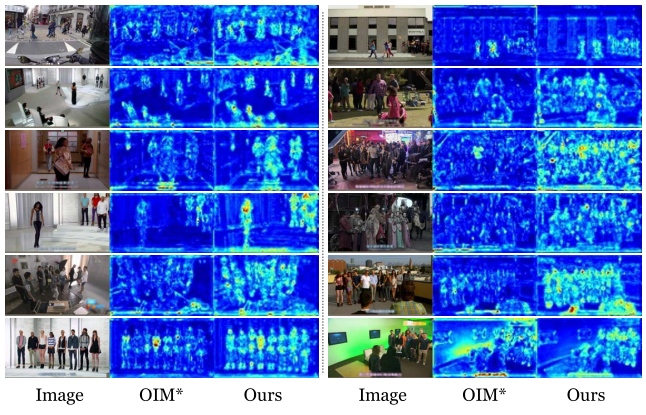


Fig. 9. Visualization of activation map of OIM* and our method. For each image triplet, the left image is the original image, the middle one is the activation map from OIM* and the one in the right is the activation map of our method.

TABLE VI

SPEED COMPARISON ON DIFFERENT GPUS, RUNTIME MEASURED IN MILLISECONDS

GPU(TFLOPs)	OIM*	NAE	NAE+	CANR	CANR+
P40(11.8)	-	158	161	-	-
V100(14.1)	-	83	98	-	-
RTX2070(7.9)	117	137	-	126	162
RTX2080Ti(13.4)	93	102	-	99	118

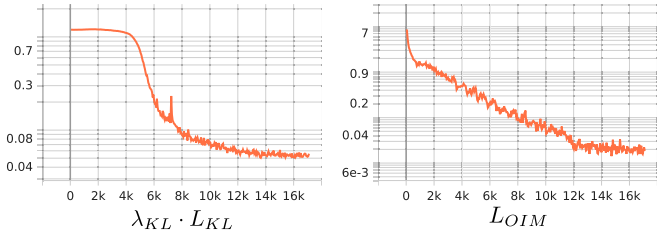


Fig. 10. Visualization of the loss curve during training on the PRW dataset.

only adds marginal overhead over OIM and is faster compared with the NAE and NAE+ while has better accuracy. With acceptable overhead, CANR+ achieves better performance.

To better understand how the data uncertainty module works during training, we visualize the loss curve of $\lambda_{KL} \cdot L_{KL}$ and L_{OIM} in Fig.10. In the early stage of the training, the re-identification feature is random and the representability is limited. Thus, we initialize the $\boldsymbol{\theta}_\sigma$ to have a small $\boldsymbol{\sigma}_i$ and facilitate the training of the re-identification features. Moreover, It's not practical to model the $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ in this stage, which leads to the plateau of L_{KL} in the early stage. After the identification feature obtain enough representability

TABLE VII
COMPARISON WITH OUR PREVIOUS WORK

Method	Backbone	mAP(%)	Top-1(%)
CANR	ResNet-50	43.41	83.81
IIDFC[58]		39.81	81.92
CANR	Res2Net-50	44.74	83.62
IIDFC[58]		43.43	83.37
CANR	ResNeSt-50	47.93	85.22

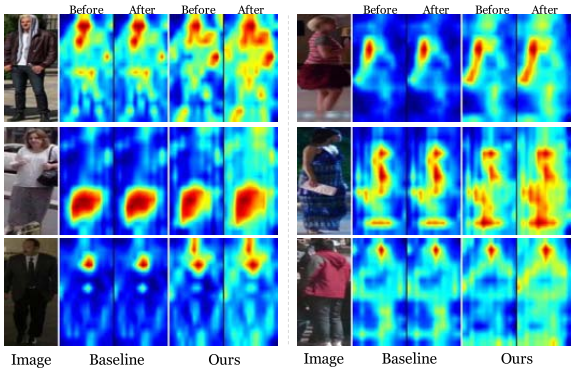


Fig. 11. Comparison of activation maps before and after the spatial attention.

(i.e. the L_{OIM} (re-identification loss) gets good enough), the L_{KL} starts to takes effect as discussed in Section-III-F.3.

We also compare CANR with our previous work [58] in Table.VII. In a fair comparison, this work shows superior performance compared with the previous work. It is also worth mentioning that this work is much simpler and does not include CIoU, HOIM, specially selected anchors, stronger backbone, etc. The performance of our work is further improved when using a stronger backbone.

E. Visualization

To have a better understanding of different methods, retrieval results of our method, the OIM, and the NAE [9] are visualized. The results are shown in Fig.8. Persons in blue bounding boxes indicate persons to be queried and the retrieval results are marked in green and red, indicating correct matches and incorrect matches respectively. For simplicity, only Top-1 retrieval results are shown. Fig.8(a), Fig.8(b), Fig.8(c), Fig.8(d) demonstrate scenes with extreme illumination condition, dramatic illumination changes, target pedestrian occluded, query pedestrian occluded respectively. CANR successfully retrieves the correct pedestrian while the NAE and OIM fail. From the results, we may conclude that CANR is more robust to illumination changes, illumination variation, and occlusion, which shows the effectiveness of CANR.

To have a deeper insight into the improvement, we visualize the activation map of OIM* and our method. The activation maps of scene images are visualized in Fig.9. Compared to the OIM*, our method tends to be more capable of extracting pedestrian features in the crowded scenario. Our method also tends to focus the discriminative patterns like shoes, textures from clothes, faces, and hairstyles. To analyze the role of spatial attention in our method, we visualize and compare

Fig. 12. Visualization of uncertainty under different level of occlusion.

the activation maps before and after the spatial attention. The results are shown in Fig.11. For each group, the first image shows the image patch corresponding to the proposal region. The latter two subgroups of images compare the visualization result of the baseline and our methods. Generally, the visualization results of the baseline are similar, which indicates that spatial attention may not work as expected in the baseline. By comparison, the spatial attention in our method tends to enhance the pedestrian regions. Moreover, our method tends to focus on more comprehensive details of pedestrians.

To have a deeper intuition on the data uncertainty. We gradually add occlusion to a pedestrian and visualize the uncertainty predicted by the network. The results are shown in Fig.12. The visualization indicates that the more occlusion added to the pedestrian, the larger data uncertainty predicted, which confirms the discussion in the Method section and demonstrates the effectiveness of the data uncertainty module.

V. CONCLUSION

In this paper, we focus on the noise in person search. We categorize the noise into scene inherent noise and human-introduced noise. For scene inherent noise, we propose a context contrastive loss and introduce the data uncertainty learning into person search. The context contrastive loss compares pedestrians in scene images to ensure the discrimination between congested pedestrians and more robust features can be obtained. As an extra advantage, the unlabeled pedestrians are fully utilized in the training process. To the best of our knowledge, we are the first to utilize the context information of scene images as supervision during the training process. The data uncertainty learning model the embedding as a distribution for more discriminative features. For scene-inherent noise, we propose a noise discovery and noise resistant mechanism to discover mislabeled pedestrians and suppress the negative influence of noise. Extensive experiments show that CANR and CANR+ can achieve competitive performance compared with state-of-the-art methods.

REFERENCES

[1] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," 2016, *arXiv:1610.02984*.

[2] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 3–19, 2013.

[3] R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Ten years of pedestrian detection, what have we learned?" in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2014, pp. 613–627.

[4] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[5] Y. Tang, B. Li, M. Liu, B. Chen, Y. Wang, and W. Ouyang, "AutoPedestrian: An automatic data augmentation and loss function search scheme for pedestrian detection," *IEEE Trans. Image Process.*, vol. 30, pp. 8483–8496, 2021.

[6] Y. Xu, B. Ma, R. Huang, and L. Lin, "Person search in a scene by jointly modeling people commonness and person uniqueness," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 937–940.

[7] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3415–3424.

[8] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1367–1376.

[9] D. Chen, S. Zhang, J. Yang, and B. Schiele, "Norm-aware embedding for efficient person search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 12615–12624.

[10] D. Chen, S. Zhang, W. Ouyang, J. Yang, and B. Schiele, "Hierarchical online instance matching for person search," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 10518–10525.

[11] W. Dong, Z. Zhang, C. Song, and T. Tan, "Bi-directional interaction network for person search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2839–2848.

[12] B. Munjal, S. Amin, F. Tombari, and F. Galasso, "Query-guided end-to-end person search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 811–820.

[13] Y. Zhong, X. Wang, and S. Zhang, "Robust partial matching for person search in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6827–6835.

[14] K. Tian, H. Huang, Y. Ye, S. Li, J. Lin, and G. Huang, "End-to-end thorough body perception for person search," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 12079–12086.

[15] X. Lan, X. Zhu, and S. Gong, "Person search by multi-scale matching," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 536–552.

[16] C. Han *et al.*, "Re-ID driven localization refinement for person search," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9814–9823.

[17] W. Dong, Z. Zhang, C. Song, and T. Tan, "Instance guided proposal network for person search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2585–2594.

[18] C. Wang, B. Ma, H. Chang, S. Shan, and X. Chen, "TCTS: A task-consistent two-stage framework for person search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11952–11961.

[19] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai, "Person search via a mask-guided two-stream CNN model," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 734–750.

[20] J. Liu, Z.-J. Zha, R. Hong, M. Wang, and Y. Zhang, "Dual context-aware refinement network for person search," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3450–3459.

[21] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu, and X. Yang, "Learning context graph for person search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2158–2167.

[22] H. Yao and C. Xu, "Joint person objectness and repulsion for person search," *IEEE Trans. Image Process.*, vol. 30, pp. 685–696, 2021.

[23] J. Chang, Z. Lan, C. Cheng, and Y. Wei, "Data uncertainty learning in face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 5710–5719.

[24] W. Yang, D. Li, X. Chen, and K. Huang, "Bottom-up foreground-aware feature fusion for person search," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3404–3412.

[25] L. Zhang, Z. He, Y. Yang, L. Wang, and X. Gao, "Tasks integrated networks: Joint detection and retrieval for image search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 456–473, Jan. 2022.

[26] Y. Zhang, X. Li, and Z. Zhang, "Efficient person search via expert-guided knowledge distillation," *IEEE Trans. Cybern.*, vol. 51, no. 10, pp. 5093–5104, Oct. 2021.

[27] X. Chang, P.-Y. Huang, Y.-D. Shen, X. Liang, Y. Yang, and A. G. Hauptmann, "RCAA: Relational context-aware agents for person search," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 84–100.

[28] H. Liu *et al.*, "Neural person search machines," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 493–501.

[29] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. ECCV*, Sep. 2018, pp. 480–496.

[30] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 274–282.

[31] H. Luo *et al.*, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2597–2609, Dec. 2020.

[32] H. Luo, W. Jiang, X. Zhang, X. Fan, J. Qian, and C. Zhang, "Alignedreid++: Dynamically matching local information for person re-identification," *Pattern Recognit.*, vol. 94, pp. 53–61, Oct. 2019.

[33] C. Zhao, X. Lv, Z. Zhang, W. Zuo, J. Wu, and D. Miao, "Deep fusion feature representation learning with hard mining center-triplet loss for person re-identification," *IEEE Trans. Multimedia*, vol. 22, no. 12, pp. 3180–3195, Dec. 2020.

[34] J. Liu *et al.*, "Multi-scale triplet CNN for person re-identification," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 192–196.

[35] H. Liu *et al.*, "Video-based person re-identification with accumulative motion context," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2788–2802, Oct. 2018.

[36] N. McLaughlin, J. M. D. Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Jun. 2016, pp. 1325–1334.

[37] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3037–3045, Oct. 2018.

[38] X. Wang, M. Liu, D. S. Raychaudhuri, S. Paul, Y. Wang, and A. K. Roy-Chowdhury, "Learning person re-identification models from videos with weak supervision," *IEEE Trans. Image Process.*, vol. 30, pp. 3017–3028, 2021.

[39] X. Wang, R. Panda, M. Liu, Y. Wang, and A. K. Roy-Chowdhury, "Exploiting global camera network constraints for unsupervised video person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 4020–4030, Oct. 2021.

[40] X. Wang, S. Li, M. Liu, Y. Wang, and A. K. Roy-Chowdhury, "Multi-expert adversarial attack detection in person re-identification using context inconsistency," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15097–15107.

[41] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 79–88.

[42] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 994–1003.

[43] X. Qian *et al.*, "Pose-normalized image generation for person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 650–667.

[44] V. Kantorov, M. Oquab, M. Cho, and I. Laptev, "ContextLocNet: Context-aware deep network models for weakly supervised localization," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2016, pp. 350–365.

[45] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3588–3597.

[46] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 768–783.

[47] J. Shi, J. Xu, B. Gong, and C. Xu, "Not all frames are equal: Weakly-supervised video grounding with contextual similarity and visual clustering losses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10444–10452.

[48] X. Zhang, Q. Chen, R. Ng, and V. Koltun, "Zoom to learn, learn to zoom," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3762–3770.

[49] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang, "Learning from massive noisy labeled data for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2691–2699.

[50] J. Gao, J. Wang, S. Dai, L.-J. Li, and R. Nevatia, "NOTE-RCNN: NOise tolerant ensemble RCNN for semi-supervised object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9508–9517.
[51] X. Wang, S. Wang, H. Shi, J. Wang, and T. Mei, "Co-mining: Deep face recognition with noisy labels," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9358–9367.
[52] A. Wang, Y. Sun, A. Kortylewski, and A. Yuille, "Robust object detection under occlusion with context-aware CompositionalNets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12645–12654.
[53] J. U. Kim, J. Kwon, H. G. Kim, and Y. M. Ro, "BBC Net: Bounding-box critic network for occlusion-robust object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 1037–1050, Apr. 2020.
[54] K. Zheng, C. Lan, W. Zeng, Z. Zhang, and Z.-J. Zha, "Exploiting sample uncertainty for domain adaptive person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 4, 2021, pp. 3538–3546.
[55] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.
[56] T. Yu, D. Li, Y. Yang, T. Hospedales, and T. Xiang, "Robust person re-identification by modelling feature uncertainty," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 552–561.
[57] D. Yu, H. Wang, P. Chen, and Z. Wei, "Mixed pooling for convolutional neural networks," in *Proc. Int. Conf. Rough Sets Knowl. Technol.*, Cham, Switzerland: Springer, 2014, pp. 364–375.
[58] S. Hou, C. Zhao, Z. Chen, J. Wu, Z. Wei, and D. Miao, "Improved instance discrimination and feature compactness for end-to-end person search," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2079–2090, Apr. 2022.
[59] C. Han, Z. Zheng, C. Gao, N. Sang, and Y. Yang, "Decoupled and memory-reinforced networks: Towards effective feature learning for one-step person search," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 2, 2021, pp. 1505–1512.
[60] X. Zhang, X. Wang, J.-W. Bian, C. Shen, and M. You, "Diverse knowledge distillation for end-to-end person search," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 4, 2021, pp. 3412–3420.
[61] Y. Yan *et al.*, "Anchor-free person search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7690–7699.
[62] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," 2015, *arXiv:1506.01497*.
[63] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. ICCV*, Oct. 2017, pp. 2961–2969.
[64] X. Zhang, R. Zhao, Y. Qiao, X. Wang, and H. Li, "AdaCos: Adaptively scaling cosine logits for effectively learning deep face representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10823–10832.

Cairong Zhao received the B.Sc. degree from Jilin University, Changchun, China, in 2003, the M.Sc. degree from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Beijing, China, in 2006, and the Ph.D. degree from the Nanjing University of Science and Technology, Nanjing, China, in 2011. He is currently a Professor with Tongji University, Shanghai, China. He is the author of more than 30 scientific articles in pattern recognition, computer vision, and related areas. His research interests include computer vision, pattern recognition, and visual surveillance.

Zhicheng Chen received the B.E. degree in computer science from Shanghai University in 2019. He is currently pursuing the master's degree with Tongji University. He was also recommended as a master's student for admission to Tongji University in 2019. His main research interests include person search, object detection, and person re-identification.

Shuguang Dou is currently pursuing the Ph.D. degree with the College of Electronics and Information Engineering, Tongji University, Shanghai, China. His research interests include computer vision, deep learning, X-ray, and person re-identification.

Zefan Qu is currently pursuing the master's degree with the College of Electronics and Information Engineering, Tongji University, Shanghai, China. His research interests include computer vision, deep learning, and person re-identification.

Jiawei Yao received the B.S. degree from Nanjing Tech University in 2010 and the M.S. and Ph.D. degrees from the University of Nottingham, U.K., in 2012 and 2016, respectively. He is currently an Associate Professor with Tongji University. He is the author of more than 20 scientific papers in environment simulation, space optimization, and related areas. His research interests include generative urban design and automation optimization.

Jun Wu (Senior Member, IEEE) received the B.Sc. degree in information engineering and the M.Sc. degree in communication and electronic systems from Xidian University, Xi'an, China, in 1993 and 1996, respectively, and the Ph.D. degree in signal and information processing from the Beijing University of Posts and Telecommunications, Beijing, China, in 1999. He joined Tongji University, Shanghai, China, as a Professor in 2010, where he is currently a Professor with the Department of Computer Science and Technology. Before he joined Tongji University, he was the Principal Scientist of Huawei and Broadcom. His research interests include wireless communication, information theory, machine learning, and signal processing.

Duoqian Miao was born in 1964. He is a Professor and a Ph.D. Tutor with the College of Electronics and Information Engineering, Tongji University, Shanghai, China. He serves as the Vice President for the International Rough Set Society, the Executive Manager for the Chinese Association for Artificial Intelligence, the Chair for the CAAI Granular Computing Knowledge Discovery Technical Committee, a Distinguished Member for the Chinese Computer Federation, the Vice President for the Shanghai Computer Federation, and the Vice President for the Shanghai Association for Artificial Intelligence. He serves as an Associate Editor for the *International Journal of Approximate Reasoning* and an Editor for the *Journal of Computer Research and Development* (in Chinese).