Dual Context-Aware Refinement Network for Person Search

Jiawei Liu[1], Zheng-Jun Zha[1*], Richang Hong[2], Meng Wang[2], Yongdong Zhang[1]
[1]University of Science and Technology of China
[2]Hefei University of Technology
{jwliu6,zhazj,zhyd73}@ustc.edu.cn
hongrc@hfut.edu.cn,eric.mengwang@gmail.com

ABSTRACT

Person search has recently gained increasing attention as the novel task of localizing and identifying a target pedestrian from a gallery of non-cropped scene images. Its performance depends on accurate person detection and re-identification simultaneously by learning effective representations. In this work, we propose a novel dual context-aware refinement network (DCRNet) for person search, which jointly explores two kinds of contexts including intra-instance context and inter-instance context to learn discriminative representation. Specifically, an intra-instance context module is designed to refine the representation for the bounding box of a pedestrian by leveraging its surrounding regions covering the same pedestrian and its accessories, which contain abundant complementary visual appearance of pedestrians. Moreover, an inter-instance context module is proposed to expand the instance-level feature for the bounding box of a pedestrian, by utilizing the rich scene contexts of neighboring co-travelers across images. These two modules are built on top of a joint detection and feature learning framework, *i.e.*, Faster R-CNN. Extensive experimental results on two challenging datasets have demonstrated the effectiveness of DCRNet with significant performance improvements over state-of-the-art methods.

CCS CONCEPTS

• **Information systems** → **Top-k retrieval in databases**.

KEYWORDS

Person search, intra-instance context, inter-instance context

ACM Reference Format:
Jiawei Liu[1], Zheng-Jun Zha[1*], Richang Hong[2], Meng Wang[2], Yongdong Zhang[1]. 2020. Dual Context-Aware Refinement Network for Person Search. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394171.3413878

Figure 1: Illustration of person re-identification and person search. Compared to person re-identification, person search is closer to real-world applications and more challenging.

1 INTRODUCTION

Person search has recently emerged as the task of localizing a specific pedestrian matching the provided query from a large-scale gallery of non-cropped scene images [2, 3, 49, 50, 52]. It requires to address pedestrian detection and re-identification simultaneously. Person search has attracted increasing attention because of its importance for many practical applications, such as automated surveillance, activity analysis and content-based retrieval [19, 19, 39, 42] *etc.* It is a challenge task due to various challenges, including cluttered background, occlusion, dramatic variations in illumination, body pose, scale and camera viewpoint, as well as similar appearance among different pedestrians *etc.*

Previous approaches [44, 49, 51, 56] for person search only focus on re-identification, which matches manually cropped pedestrian images across different cameras. Although these approaches have achieved steady performance improvement on popular benchmarks, they have major restriction for practical usage, due to the assumption of precise pedestrian detection. In real world application, perfect pedestrian bounding boxes are either unavailable or expensive to obtain. Moreover, the off-the-shelf pedestrian detectors would unavoidably produce misalignments and misdetections, thus resulting in compromising the final performance. Figure 1 illustrates the comparison between person re-identification and person search.

* Corresponding author.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413878

To close this gap, state-of-the-art methods [3, 16, 46, 48] unify pedestrian detection and re-identification into an end-to-end framework, which mainly based on Faster R-CNN [41] detection model. Specifically, an auxiliary fully-connected layer is added on the top convolutional layer of Faster R-CNN to extract discriminative visual features for re-identification. During the training stage, they optimize a joint loss, which is composed of a Faster R-CNN detection loss and a well-designed identification loss. However, these methods still employ individual representation for person localization and matching. Therefore, it is difficult to search pedestrians with large intra-class variation and small inter-class variation, especially in situations where require to retrieve through a huge gallery set.

Many empirical studies [4, 5, 8, 12, 33] have suggested that the performance can be greatly improved by proper modeling of context in recognition tasks. Thus, a promising solution for addressing the above issue in the person search task, is to utilize the rich contextual information contained in the scene images. We observe that for a bounding box of a pedestrian, its surrounding regions covering the same pedestrian and its accessories usually contain useful contextual information on pedestrian appearance, *e.g.*, hat, bag and bike (Intra-instance context), which is beneficial for pedestrian localization and re-identification. Moreover, pedestrians often tend to walk alongside others or in a group [1, 27, 34]. Other neighboring pedestrians appearing in the same scene thus contain rich contexts (Inter-instance context), which can reduce the ambiguity significantly in retrieving pedestrians that are partially occluded or have similar appearance with other pedestrians [45, 57]. These two kinds of contexts provide different types of complementary information, which can be combined together to jointly help searching pedestrians. A few of works [2, 28, 36] make a preliminary effort on learning representation with contextual information, which simply utilize the contexts mixed with various information in the whole scene images, even some contexts are noise information (*e.g.*, the other neighboring pedestrians only appear once in the scene images). The method [50] further employs neighboring co-travelers in the scene images as useful contextual information to improve the performance. Nevertheless, it dose not sufficiently encode these inter-instance contexts by treating them equally important, and neglects the crucial intra-instance contexts.

In this work, we propose a novel dual context-aware refinement network (DCRNet) to explore contextual information for person search. It jointly employs two kinds of contexts including intra-instance context and inter-instance context to learn more effectiveness representation for pedestrian localization and re-identification. As illustrated in Figure 2, DCRNet consists of a Faster R-CNN with an intra-instance context module and an inter-instance context module. Given a pair of query-gallery images, DCRNet firstly generates bounding box candidates of pedestrians and extracts low-level features from them. As these low-level features are not powerful enough to identify different pedestrians, the two well-deigned modules are used to refine the features with the contextual information. Specifically, for each bounding box of a pedestrian, the intra-instance context module identifies its surrounding regions covering the same pedestrian and its accessories, and aggregates the contextual information of the surrounding regions into a unified contextual representation based on an adaptive weighting strategy. The module finally utilizes a squeeze-and-excitation block (SEBlock) [17] to refine the low-level features with the unified contextual representations. Moreover, for the target pedestrian, the inter-instance context module filters the neighboring pedestrians appearing in the query and gallery images as informative inter-instance context. A context graph model is built with a Gated Recurrent Unit (GRU) [6] to encode the representation of the target pedestrian with the inter-instance contextual information for generating a more meaningful representation, which is then used to calculate the global similarity of query-gallery pair. With the two modules, DCRNet is able to learn effective representations of pedestrians, leading to satisfactory person search results. We conduct extensive experiments to evaluate DCRNet on the challenging benchmarks, and report superior performance over state-of-the-art approaches.

The main contribution of this paper is three-fold: (1) We propose a dual context-aware refinement network (DCRNet) to explore contextual information for person search; (2) We design an intra-instance context module to refine the representations of pedestrians by leveraging the useful contextual cues of the surrounding context regions; (3) We design an inter-instance context module to expand the instance-level representations of pedestrians by utilizing the scene contextual cues of neighboring co-travelers across images.

2 RELATED WORK

In this section, we first review existing works on person search, which is composed of two stages: pedestrian detection and person re-identification. Hence, we also review some recent works in both fields.

Person Search. Person search is a recently developed task [2, 3, 36, 50], which aims to match a target pedestrian from a great number of whole scene images. In the literature, there are two approaches to deal with this task. Some methods [3, 16, 21, 46, 48] jointly train pedestrian detection and re-identification model in a single framework without the guidance of query image. Xiao *et al.* [48] jointly handled pedestrian detection and person re-identification in a single convolutional neural network. Meanwhile, An Online Instance Matching (OIM) loss function was designed, which was scalable to datasets with numerous identities. Xiao *et al.* [46] proposed an Individual Aggregation Network (IAN) that can accurately localize pedestrians by learning to minimize intra-person feature variations with a center loss. Other methods [2, 28, 36, 50] focus on designing query-guided models for person search in an end-to-end manner. Munjal *et al.* [36] proposed a query-guided end-to-end person search network (QEEPS), which utilized the query for its global context and local cues. Yan *et al.* [50] proposed a contextual instance expansion module and a graph learning module, which were employed to search and filter useful context information in the query and gallery scene images, and aggregate the contextual information to update the target similarity.

Person Re-Identification. Recently deep learning dominates person re-identification research community with significant advantages in retrieval performance [29, 30, 32, 35, 47]. Most methods aim at extracting robust and discriminative representations and/or learning appropriate distance metric with customized loss functions [15, 18, 22, 23]. Zhao *et al.* [54] proposed an algorithm of feature disentangling and temporal aggregation for video-based person re-identification, in which an attribute-driven method was proposed

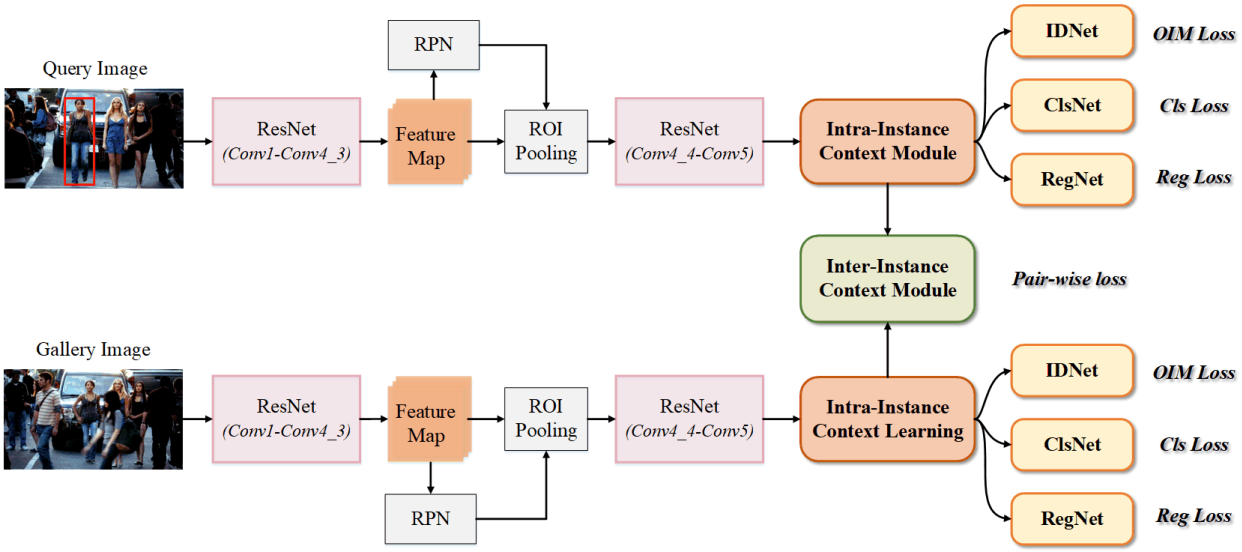

Figure 2: The overall architecture of the proposed DCRNet. It consists of a Faster RCNN with an intra-instance context module and an inter-instance context module. The two modules are utilized to refine the representations of the detected bounding box candidates with two kinds of contexts for accurate pedestrian detection and re-identification.

for feature disentangling and a transfer learning method was introduced for automatically annotating attribute labels. Shen *et al.* [43] proposed a similarity-guided graph neural network to incorporate the rich gallery-gallery similarity information into training process of person re-identification. Liu *et al.* [31] proposed a deep adversarial graph attention convolution network (A-GANet) for text-based person re-identification, which exploited both textual and visual scene graphs, towards learning fine-grained structural textual and visual representations.

Pedestrian Detection. Early works on pedestrian detection were built upon hand-crafted features and linear classifiers. Typical methods include Deformable Part Model (DPM) [11], Aggregated Channel Features (ACF) [9] and Integrate Channel Features (ICF) [37]. Recently, many deep learning based methods have been proposed to boost the performance of pedestrian detection, which can be roughly divided into one stage methods [10, 26, 40] and two-stage methods [7, 13, 25, 41]. The main difference between them is whether to generate the proposals. For one stage methods, Duan *et al.* [10] proposed a representative one-stage keypoint-based detector named CornerNet with cascade corner pooling and center pooling, which detected each object as a triplet, rather than a pair, of keypoints, and improved both precision and recall. For two-stage methods, the prominent representative work is the Faster R-CNN [41], which proposed a region proposal network (RPN). It greatly reduces the amount of computation while shares the characteristics of the backbone network.

3 METHOD

In this section, we firstly present the overall architecture of the proposed DCRNet, and then introduce each component of the architecture in the following subsections.

3.1 Overall Architecture

Although deep learning models have greatly improved the representation ability of individual pedestrian features, it is still difficult to search the target pedestrian under various complicated situations. Therefore, we propose a novel dual context-aware refinement network for person search that explores two kinds of rich contextual information including intra-instance context and inter-instance context to learn better representation. As shown in Figure 2, DCRNet consists of a Faster R-CNN with an intra-instance context module and an inter-instance context module. Faster R-CNN is based on a ResNet-50 backbone [14], which contains five residual blocks (named Conv1_x to Conv5_x), region proposal net (RPN) and Region of Interest (ROI) pooling layer.

Given a pair of query-gallery images, DCRNet uses the first four residual blocks (Conv1 to Conv4_3) to transform raw pixels into 1024 channel feature maps, which have 1/16 resolutions of the input images. On top of these feature maps, RPN is built to produce pedestrian proposals, which is further passed to a $512 \times 3 \times 3$ convolutional layer to transform the features specifically for pedestrians. Similar to previous methods [36, 41, 48], we associate 9 anchors at each feature map location. Two losses are utilized to optimize the RPN, *i.e.*, a binary classification loss to judge whether the anchor is a person or not, and a regression loss to perform bounding box regressions. After that, non-maximum suppression algorithm [38] is utilized to remove duplicated detections and 128 candidate proposals are selected for each image. All the candidate proposals are fed into the RoI Pooling layer to generate fixed $1024 \times 14 \times 14$ regions from their feature maps. Then, they are passed through the rest Conv4_4 to Conv5_3 of ResNet-50 model, connected with a global average pooling layer to generate initial 2048 dimensional features.

In order to improve the capacity of the extracted features, these bounding box proposals are passed through the intra-instance context module and inter-instance context module for feature refinement with contextual information. For each bounding box, the intra-instance context module firstly identifies its surrounding context regions covering the same pedestrian from all the bounding boxes. The contextual information of these surrounding context regions are gathered by an adaptive weighting strategy into a unified contextual representation, which is then used to refine the initial feature by a SEBlock. Followed with the work [36], the preliminary refined features are fed into a IDNet with OIM loss, a ClsNet with binary classification loss and a RegNet with regression loss for distinguishing different pedestrians, making person/non-person judgments and further adjusting the locations, respectively. These three networks are implemented by fully connected (FC) layers. In addition, the preliminary refined features are fed into the inter-instance context module for further improvement by utilizing useful contextual information of neighboring co-travelers across images. For the bounding boxes of the target pedestrian, the inter-instance context module filters the neighboring pedestrians appearing in both the query and gallery images as informative inter-instance context. A context graph model is designed with a GRU to further encode the inter-instance contextual information, and transfer the preliminary refined features into more meaningful ones, which are finally utilized to model the global similarity of query-gallery pair with a pair-wise loss. The calculated similar scores are applied for inference, during the testing stage.

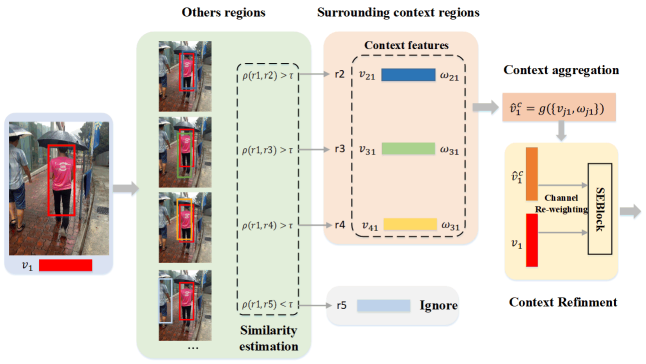

Figure 3: Detailed structure of the intra-instance context module. It is used to refine the representations of pedestrians by leveraging the useful contextual cues of the surrounding context regions.

3.2 Intra-Instance Context Module

The intra-instance context module is proposed to refine the feature for the bounding box of a pedestrian by leveraging its surrounding regions covering the same pedestrian and its accessories, which contain rich contextual information on visual appearance of pedestrians. The intra-instance context can deliver informative clues for describing the accurate status of a pedestrian. Hence, the module is able to help the framework generate more reliable bounding boxes and produce more discriminative embeddings for person search, which is shown in Figure 3.

Specifically, DCRNet takes a pair of query-gallery images as input. After each image going through the global average pooling layer, the network generates 128 bounding box proposals and produces the initial 2048 dimensional features $\boldsymbol{v}$. For each bounding box $\boldsymbol{r}_i$, the intra-instance context module firstly identifies its surrounding context regions from all bounding boxes for collecting rich contextual information. The surrounding context regions represent closely related regions covering the same pedestrians with the selected bounding box. In order to obtain an adequate set of surrounding context regions, we estimate the closeness between the selected bounding box and other bounding boxes, and define the closeness between two bounding boxes by the concept of correlation level $\rho(\boldsymbol{r}_i, \boldsymbol{r}_j)$. The set of surrounding context regions for $\boldsymbol{r}_i$ is described as: $\boldsymbol{R}_i^c = \{\boldsymbol{r}_j | \rho(\boldsymbol{r}_i, \boldsymbol{r}_j) > \tau\}$, where τ refers to a threshold. The correlation level of two bounding boxes is calculated by their Intersect-over-Union (IoU) score $\rho(\boldsymbol{r}_i, \boldsymbol{r}_j) = IoU(\boldsymbol{r}_i, \boldsymbol{r}_j)$. $\boldsymbol{R}_i^c$ contains abundant complementary information that is beneficial for rectifying the coordinates of bounding box and improving the discriminability of the extracted features.

Considering that the number of surrounding context regions for different bounding boxes is not fixed and could range from zero to hundreds, it is different to conduct appropriate feature refinement by using an arbitrary amount of contextual information. Thus, we introduce an aggregation function to fuse all the contextual information of $\boldsymbol{R}_i^c$ into a unified contextual representation $\hat{\boldsymbol{v}}_i^c$ by an adaptive weighting strategy. For the selected bounding box $\boldsymbol{r}_i$, the contextual information carried by a surrounding context region $\boldsymbol{r}_j$ is represented as its feature $\boldsymbol{v}_{ji}$. Different surrounding context regions contains different levels of beneficial information for the selected bounding box. The more related regions should make major contributions to the unified contextual representation. We refer ω_{ji} as the weight score of $\boldsymbol{v}_{ji}$, which is computed as follows:

$$\omega_{ji} = \omega_{ji}^s \cdot \omega_{ji}^p \tag{1}$$

where ω_{ji}^s refers to the semantic relation between the two bounding boxes, ω_{ji}^p represents the spatial position relation between the two bounding boxes. They are defined as follows:

$$\begin{cases} \omega_{ji}^s = cosine(\boldsymbol{v}_i, \boldsymbol{v}_{ji}) \\ \omega_{ji}^g = IoU(\boldsymbol{r}_i, \boldsymbol{r}_j) \end{cases} \tag{2}$$

Before calculating the semantic relation, the extracted features are applied with L2-normalization operation. The aggregation function is formulated as follows:

$$\hat{\boldsymbol{v}}_i^c = \frac{\sum_j \omega_{ji} \cdot \boldsymbol{v}_{ji}^c}{\sum_j \omega_{ji}}, \quad \boldsymbol{r}_j \in \boldsymbol{R}_i^c \tag{3}$$

Finally, we utilize the squeeze-and-excitation (SE) technique for refining the feature $\boldsymbol{v}_i$ of the select bounding box with the unified contextual representation, which is formulated as follows:

$$\begin{aligned} \boldsymbol{v}_i &= \boldsymbol{v}_i + \boldsymbol{s} \odot \hat{\boldsymbol{v}}_i^c \\ \boldsymbol{s} &= \sigma(\boldsymbol{W}_2 \delta(\boldsymbol{W}_1 [\boldsymbol{v}_i, \hat{\boldsymbol{v}}_i^c])) \end{aligned} \tag{4}$$

where σ and δ refer to the Sigmoid activation and the ReLU function [20], respectively. $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are the weight parameters of two

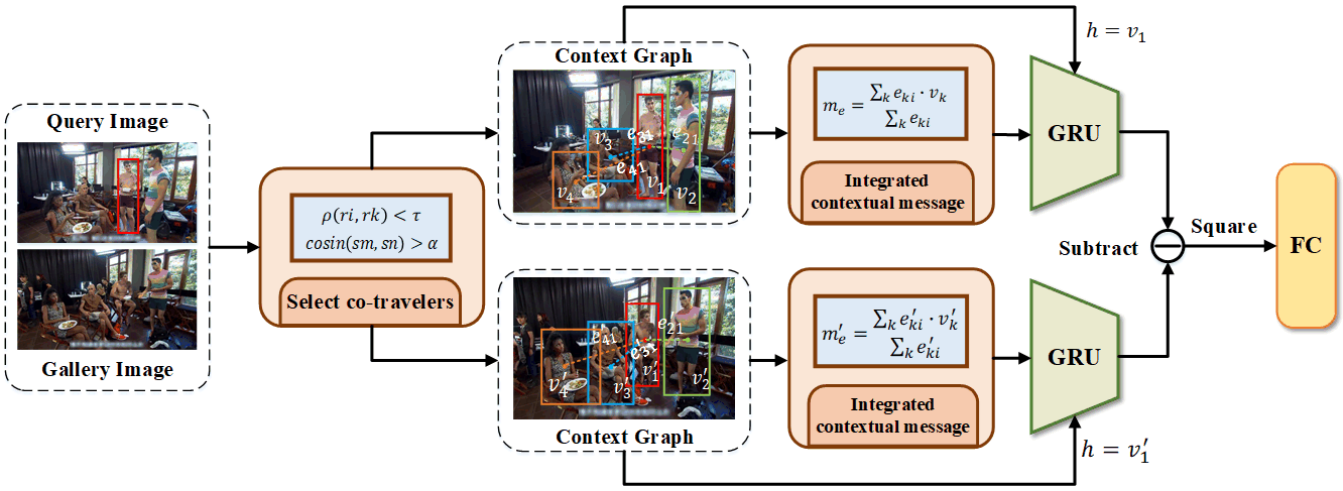

Figure 4: Detailed structure of the inter-instance context module. It is used to expand the instance-level representations of pedestrians by utilizing the scene contextual cues of neighboring co-travelers across images.

FC layers. The SEBlock learns to selectively emphasis beneficial context and suppress less useful ones by re-weighting channel features. The refined feature is then connected with IDNet, ClsNet and RegNet respectively for pedestrian detection and re-identification. In addition, it is fed into the inter-instance module for further feature refinement and calculating the final global similarity of the query-gallery pair.

3.3 Inter-Instance Context Module

The inter-instance context module is proposed to expand the instance-level feature for the bounding box of a pedestrian, by using the informative scene contexts of neighboring co-traveler across images. As pedestrians are likely to walk alongside others or in groups, the neighboring pedestrians appearing in the same scene contain crucial context cues, which can provide complementary information for assisting to search the target person from great variations. Figure 4 illustrates the detailed structure of the inter-instance context module.

Given an bounding box $\boldsymbol{r}_i$ of the target pedestrian in the query image, the inter-instance context module firstly filters the bounding boxes of neighboring context pedestrians that appear in both the query and gallery images, from all the bounding boxes. Specifically, we pick out the bounding boxes $\boldsymbol{R}_i^d = \{\boldsymbol{r}_k | \rho(\boldsymbol{r}_i, \boldsymbol{r}_k) < \tau\}$ belonging to other pedestrians in the query image by computing the correlation level. After that, we compute the similarity scores between all the selected bounding boxes in the query images I_q and all bounding boxes in the gallery image I_g, *i.e.*, $s(m, n) = cosine(\boldsymbol{v}_m, \boldsymbol{v}_n)$, $\boldsymbol{r}_m \in I_q$, $\boldsymbol{r}_n \in I_g$, and set a threshold α to make binary decision whether the other neighboring pedestrians in the query image appear in the gallery image. With the collected inter-instance contexts, we propose a context graph model with a GRU to make full use of these information and make more confident judgment whether the target pair across the two images belongs to the same person ID. Two context graphs are structured for the pair of query-gallery images. Supposing a context graph $\mathcal{G}_v = (O_v, \mathcal{E}_v)$ consisting of N nodes O and a set of edges $\mathcal{E}$. The nodes correspond to the features of the selected bounding boxes of pedestrians appearing in both the two images, which consist of a target pedestrian and K neighboring context pedestrians ($N = K + 1$). In this context graph, the node of the target pedestrian is the center of the context graph, which is connected to all other nodes of the neighboring context pedestrians for information propagation and feature refinement.

For the target node, the key of interaction is to encode the contextual messages passed from the other nodes. Considering that the target node requires to receive multiple incoming contextual messages, it is necessary to build an aggregation function, which can memory the node details itself and then fuse incoming contextual messages from other nodes into a more meaningful feature. As this aggregation function behaves like a memory model, we utilize a GRU, which can map from the whole history of previous inputs to each output. The memory cell of GRU allows the hidden state to drop any information that is found to be unrelated with input later through the reset gate, and controls how much information from the previous state will propagate to the current hidden state through the update gate. We take the feature of the target node as the initial state of the GRU, and treat the incoming contextual messages from other nodes as input. Since the target node is connected to all other nodes, we require to calculate an integrated contextual message $\boldsymbol{m}_e$ in advance, which is then taken as the input of the GRU. Different nodes provide different contributions to the target node, thus each node-node relationship e_{ki} is computed as a scalar weight, which represents the influence of $\boldsymbol{v}_k$ on $\boldsymbol{v}_i$. The relationship between nodes is common determined by spatial position relationship and visual relationship. The integrated contextual message to node $\boldsymbol{v}_i$ is calculated by:

$$\begin{aligned} \boldsymbol{m}_e &= \frac{\sum_k e_{ki} \cdot \boldsymbol{v}_k}{\sum_k e_{ki}} \\ e_{ki} &= relu(\tilde{\boldsymbol{W}}_p \boldsymbol{P}_{ki}) * tanh(\tilde{\boldsymbol{W}}_v [\boldsymbol{v}_k, \boldsymbol{v}_i]) \end{aligned} \tag{5}$$

where $\tilde{\boldsymbol{W}}_p$ and $\tilde{\boldsymbol{W}}_v$ are weight parameters. The visual relationship is obtained by concatenating the feature $\boldsymbol{v}_k$ and $\boldsymbol{v}_i$. $\boldsymbol{P}_{ki}$ represents the spatial position relationship, which is defined as follows:

$$\begin{aligned}\boldsymbol{P}_{ki} = [w_k, h_k, a_k, w_i, h_i, a_i, \frac{x_i - x_k}{w_k}, \frac{y_i - y_k}{h_k}, \\ \frac{(x_i - x_k)^2}{w_k^2}, \frac{(y_i - y_k)^2}{h_k^2} \log(\frac{w_i}{w_k}), \log(\frac{h_i}{h_k})]\end{aligned} \tag{6}$$

where (x_i, y_i) is the center coordinate of the bounding box $\boldsymbol{r}_i$, w_i, h_i are the width and height of $\boldsymbol{r}_i$, and a_i is the area of $\boldsymbol{r}_i$. When the feature of target node updated, the relationships between nodes will change at the next timestep. GRU would take the updated integrated contextual message as new input, then compute the next refined feature for the target node. More timesteps of updating make the module more stable and effective. Finally, the expanded refined features of the target nodes in the two graphs are obtained. The two features are applied with element-wise subtraction and square operations, and then fed into a 2-dimensional FC layer. A pair-wise loss is connected to the FC layer, which is utilized to model the global similarity of query-gallery pair and optimize the network.

3.4 Loss Function

The typical Faster R-CNN detector [41] is supervised with loss functions for classification ($\mathcal{L}_{cls}$), regression ($\mathcal{L}_{reg}$), RPN objectness ($\mathcal{L}_{rpn_o}$), and RPN box regression ($\mathcal{L}_{rpn_r}$). Although Softmax loss (Identification loss) is wildly employed for re-identification task, it is difficult for person search task to train a Softmax layer as the number of person IDs is large. The person IDs appeared in each mini-batch are highly sparse, which makes Softmax loss learned inefficient. Moreover, this loss can not exploit the unlabeled pedestrians in the scene images, which also contain useful information for person search. To address this problem, we adopt an online instance matching (OIM) loss [48, 50] $\mathcal{L}_{cls}$ to supervise the network for pedestrian classification.

Besides, we introduce a pair-wise loss to supervise the network and model the global similarity scores of query-gallery pairs, which is formulated as follows:

$$\mathcal{L}_{pair} = -\frac{1}{M}\sum_{i=1}^{M} \log \frac{e^{\boldsymbol{W}_{p,y_i}^T \boldsymbol{z}_i + \boldsymbol{b}_p}}{\sum_{j=1}^{2} e^{\boldsymbol{W}_{p,j}^T \boldsymbol{z}_j + \boldsymbol{b}_p}} \tag{7}$$

where y_i is a 2-dim vector, indicating that the detected target pair of pedestrians in the query and gallery images whether have the same person ID. $\boldsymbol{z_i}$ refers to the output of the last FC layer in the inter-instance module. $\boldsymbol{W}_{p,j}$ represents the j-th column of the weight matrix $\boldsymbol{W}_p$. Thus, the total loss for the DCRNet is defined as follows:

$$\begin{aligned}\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{cls} + \lambda_2 \cdot \mathcal{L}_{reg} + \lambda_3 \cdot \mathcal{L}_{rpn_o} + \lambda_4 \cdot \mathcal{L}_{rpn_r} \\ + \lambda_5 \cdot \mathcal{L}_{OIM} + \lambda_6 \cdot \mathcal{L}_{pair}\end{aligned} \tag{8}$$

where λ_{1-6} denotes the balance weights of the loss terms.

4 EXPERIMENTS

In this section, we first introduce the two person search datasets, and the evaluation protocols that we use in the experiments, as well as some implementation details. After that, we show experimental results with comparison to state-of-the-art methods, followed by several ablation studies to explore the effects of different components in DCRNet.

4.1 Experimental Settings

Datasets- We evaluate DCRNet on both CUHK-SYSU [48] and PRW [56] datasets. CUHK-SYSU is a large-scale person search dataset, which is captured by a hand-held camera or chosen from movie snapshots. It contains 18,184 images and 96,143 annotated bounding boxes indicating the locations of different pedestrians. There are a total of 8,432 labeled identities, and the rest of the pedestrians are treated as negative samples for re-identification. The training set contains 11,206 images and 5,532 identities, while the testing set consists 2,900 query pedestrians and 6,978 gallery images. For each query person, there exists at least two images containing the target pedestrian in the gallery set. PRW dataset is captured on a university campus by six cameras. It consists of 11,816 images annotated with 34, 304 bounding boxes. Among all the pedestrians, 932 identities are labeled and the rest of them are marked as unknown pedestrians. The training set contains 5,134 images and 482 identities, while the testing set includes 6,112 gallery images and 2,057 query images with 450 identities.

Evaluation Metrics- We employ two commonly-used performance metrics: mean Average Precision (mAP) and Common Matching Characteristic (CMC top-K) for evaluation. mAP is derived from the object detection task and reflects the accuracy of localizing the query in all gallery images (AP is computed for each ID and averaged to compute the mAP). CMC top-K is inherited from the person re-identification task, where a matching is counted if there is at least one of the top-K predicted bounding boxes overlaps with the ground truths with intersection-over-union (IoU) greater or equal to 0.5.

Implementation Details- The implementation of the proposed method is based on the Pytorch framework with four NVIDIA Titan RTX GPUs. The input images are re-scaled such that their shorter side is 600 pixels. We pad or crop the query images to the same size of the gallery one. The scales and aspect ratios for anchor size is the same as the typical Faster R-CNN. In addition, we augment the inputs by flipping both the query and the gallery images. The number of mini-batches is 16 scene images. The learning rate is initialized to 0.01, which is reduced by a factor of 10 after 10 epochs. We train the whole framework using SGD optimizer with Nesterov momentum of 0.9 and the weight decay of $5e^{-4}$ for 20 epochs in total. The parameters τ, α are set to 0.5 and 0.6, respectively. The balance weight of λ_{1-6} are all set to 1. The setting of OIM loss is followed with the previous work [36].

4.2 Comparison to State-of-the-Arts

CUHK-PEDES: Table 1 shows the performance comparison of the proposed DCRNet against state-of-the-art methods on CUHK-SYSU dataset. The gallery size for all the methods are set to 100. The compared methods include CNN+DSIFT+Euclidean [53], CNN+DSIFT-+KissMe [53], CNN+BoW+Cosine [55], CNN+LOMO+XQDA [24], OIM [48], IAN [46], NPSM [28], I-Net [16], MGTS [3], LCG [50], QEEPS [36]. "CNN" denotes the Faster R-CNN detector based on ResNet-50 model. Compared with results from using hand-crafted features [24, 53, 55], we observe that other deep learning based

Table 1: Performance comparison to the state-of-the-art methods on CUHK-SYSU dataset.

Method	mAP(%)	top-1(%)
CNN+DSIFT+Euclidean [53]	34.5	39.4
CNN+DSIFT+KissMe [53]	47.8	53.6
CNN+BoW+Cosine [55]	56.9	62.3
CNN+LOMO+XQDA [24]	68.9	74.1
OIM [48]	75.5	78.7
IAN [46]	76.3	80.1
NPSM [28]	77.9	81.2
I-Net [16]	79.5	81.5
MGTS [3]	83.0	83.7
LCG [50]	84.1	86.5
QEEPS [36]	84.4	84.4
DCRNet	**87.5**	**88.7**

feature representation methods achieve significant performance improvements. Especially, DCRNet achieves 87.5% mAP score and 88.7% top-1 accuracy, respectively. It can seen that our method surpasses existing methods, improving the 2nd best compared method QEEPS by 3.1% mAP score and 4.3% top-1 accuracy, which demonstrates the effectiveness of the proposed method. Moreover, DCRNet also achieves significant performance improvement as compared to LCG with only utilizing inter-instance contextual information. The comparison indicates the proposed DCRNet is able to capture more effective representation by employing both the intra-instance and inter-instance contexts. An illustration of some retrieval results on this dataset is given in Figure 6.

PRW: We further compare our method with 9 state-of-the-art approaches, including LDCF+LOMO+XQDA [24], LDCF+IDE [56], LDCF+IDE+CWS [56], OIM [48], IAN [46], NPSM [28], MGTS [3], LCG [50], QEEPS [36] on PRW dataset. The results are shown in Table 2. Following the dataset setting [56], the gallery contains all the 6,112 testing images. Compared to CUHK-SYSU dataset, all the methods achieve poorer results on PRW dataset. This is mainly due to lacking sufficient training data on PRW dataset, which would limit the generalization capability of these models. It can be seen that DCRNet obtains 38.8% mAP and 77.7 % top-1 accuracy. It improves the second best result of method QEEPS by 1.7% mAP score and 1.0% top-1 accuracy. This confirms the effectiveness of our method by employing two kinds of contexts for the person search task.

4.3 Ablation Studies

To demonstrate the effectiveness and contribution of each component of the proposed DCRNet, we conduct a series of ablation experiments on CUHK-SYSU dataset. We evaluate the influence of the intra-instance context module and the inter-instance context module, and compare their performance for person search with different experimental settings.

Table 3 summarizes the ablation results of the proposed DCRNet. DCRNet w/o Intra refers to DCRNet without the intra-instance context module, which only exploits the inter-instance contexts of neighboring co-travelers for feature refinement. DCRNet w/o Inter refers to DCRNet without the inter-instance context module, which only uses the intra-instance contexts of surrounding context regions to enhance the features. From Table 3, DCRNet w/o Intra obtains 84.8% mAP score and 85.9% top-1 accuracy respectively, DCRNet w/o Inter obtains 85.2% mAP score and 86.4% top-1 accuracy respectively. By comparing DCRNet w/o Intra with DCRNet, we can observe that the intra-instance context module is able to capture the useful contextual information from surrounding context regions, which can deliver informative clues for describing the accurate status of an pedestrian. By comparing DCRNet w/o Inter with DCRNet, we can observe that the inter-instance context module can encode the rich scene contexts of neighboring co-travelers and learn more effective features to improve the performance. Moreover, DCRNet yields the best performance of 87.5% mAP score and 88.7% top-1 accuracy, which indicts that these two kinds of contexts are complementary to jointly help search the target pedestrians.

Table 4 summarizes the ablation results of the adaptive weighting strategy and the SEblock in the intra-instance context module, as well as the GRU in the inter-instance context module. DCRNet w/o Adaption refers to DCRNet replacing the adaptive weighting strategy in the intra-instance context module with an average pooling strategy. DCRNet w/o SEblock refers to DCRNet replacing the SEblock in the intra-instance context module with addition operation. DCRNet w/o GRU refers to DCRNet removing the GRU in the

Table 2: Performance comparison to the state-of-the-art methods on PRW dataset.

Method	mAP(%)	top-1(%)
LDCF+LOMO+XQDA [24]	11.0	31.1
LDCF+IDE [56]	18.3	44.6
LDCF+IDE+CWS [56]	18.3	45.5
OIM [48]	21.3	49.9
IAN [46]	23.0	61.9
NPSM [28]	24.2	53.1
MGTS [3]	32.6	72.1
LCG [50]	33.4	73.6
QEEPS [36]	37.1	76.7
DCRNet	**38.8**	**77.7**

Table 3: Evaluation of the effectiveness of each component within DCRNet on CUHK-SYSU dataset.

Model	mAP(%)	top-1(%)
DCRNet w/o Intra	84.8	85.9
DCRNet w/o Inter	85.2	86.4
DCRNet	87.5	88.7

Table 4: Evaluation of the effectiveness of each component within the intra-instance context module and the inter-instance context module on CUHK-SYSU dataset.

Model	mAP(%)	top-1(%)
DCRNet w/o Adaption	85.8	86.9
DCRNet w/o SEBlock	86.0	87.3
DCRNet w/o GRU	86.4	87.5
DCRNet	87.5	88.7

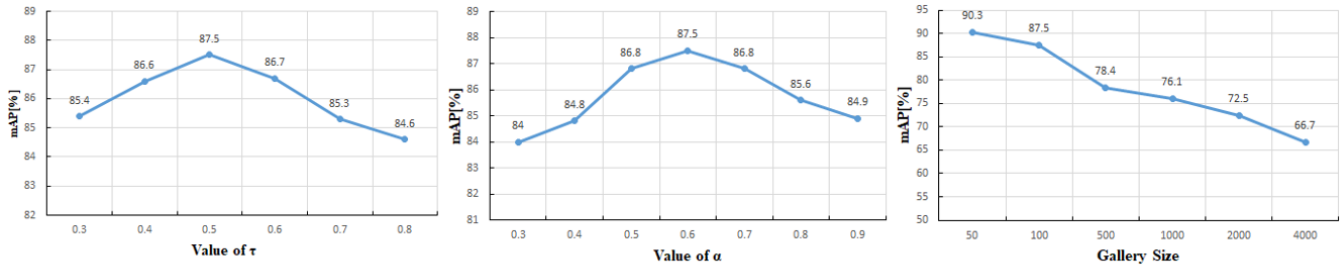

Figure 5: (a) Performance on CUHK-SYSU dataset with different values of τ. (b) Performance on CUHK-SYSU dataset with different values of α. (c) Performance on CUHK-SYSU dataset with varying gallery size.

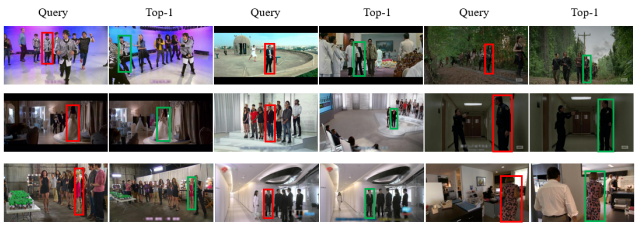

Figure 6: Qualitative Top-1 person search results for some challenging query images by the proposed DCRNet on CUHK-SYSU dataset. The red bounding boxes in the query images represent the query pedestrians, and the green ones indicate the corresponding pedestrians with correct matching in the gallery images.

inter-instance context module. From Table 4, DCRNet w/o Adaption obtains 85.8% mAP score and 86.9% top-1 accuracy, DCRNet w/o SEBlock obtains 86.0% mAP score and 87.3% top-1 accuracy, as well as DCRNet w/o GRU obtains 86.4% mAP score and 87.5% top-1 accuracy. The performance improvement of DCRNet over DCRNet w/o Adaption and DCRNet w/o SEblock indicates that the adaptive weighting strategy and the SEBlock in the intra-instance context module is able to aggregate precisely the contexts and generate effective comprehensive representation from bounding boxes of pedestrians. The performance of DCRNet w/o GRU is inferior to DCRNet, indicating that the GRU can remember long-term information, choose the useful context cues and aggregate these information into a more meaningful representation, thus benefits the performance.

We also conduct experiments to evaluate the impact of the parameters of τ and α on performance for the proposed DCRNet. The results are visualized in Figure 5(a), (b). From the results, we find that the performance will first increase as the parameters of τ and α grow. This is because of the extracted intra-instance and inter-instance contexts are highly confident ones, and thus brings useful information to improve the effectiveness of the representations. After DCRNet obtains the best performance with $\tau = 0.5$ and $\alpha = 0.6$, the performance will drop dramatically, which is due to filtering a mass of positive intra-instance and inter-instance contexts. Therefore, τ and α are set to 0.5 and 0.6, respectively for DCRNet. In order to evaluate the performance scalability of the proposed DCRNet, we compare with our methods under different gallery sizes. Figure 5(c) shows how the mAP score changes with a varying gallery size of [50, 100, 500, 1000, 2000, 4000]. We observe that when the size of gallery set increases from 50 to 4000, the declining extent of our method is relatively small, which verifies the scalability of DCRNet when dealing with large-scale person search problem. With the increasing scales of gallery set, more distracting pedestrians are involved in the searching process, which is close to practice applications. This indicates the importance of learning more effective features for the bounding boxes.

5 CONCLUSION

In this work, we propose a novel dual context-aware refinement network (DCRNet) for person search, which explores two kinds of contextual information from scene images to obtain discriminative representations of pedestrians. The proposed DCRNet learns the representation for the bounding box of a pedestrian by utilizing the intra-instance context of its surrounding regions covering the same pedestrian and its accessories, and the inter-instance context of neighboring co-travelers across images. These two kinds of contexts contain abundant complementary visual appearance of pedestrians and are able to reduce the visual ambiguity significantly in retrieving pedestrians. They can be jointly used to enhance the capacity of the representations. Moreover, DCRNet can be built on any joint detection and feature learning frameworks, such as Faster R-CNN model. We conducted extensive experiments on two widely-used person search datasets, *i.e.*, CUHK-SYSU and PRW. The experimental results have shown that the proposed DCRNet improves the performance of person search over a wide range of state-of-the-art methods.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China under Grant 2017YFB1300201, the National Natural Science Foundation of China (NSFC) under Grants U19B2038, 61620106009, 61525206, 61932009, 61722204, 61725203, and the China Postdoctoral Science Foundation Funded Project under Grant 2020M671898.

REFERENCES

[1] Min Cao, Chen Chen, Xiyuan Hu, and Silong Peng. 2017. From groups to co-traveler sets: Pair matching based person re-identification framework. In *Proceedings of the IEEE International Conference on Computer Vision*. 2573–2582.

[2] Xiaojun Chang, Po-Yao Huang, Yi-Dong Shen, Xiaodan Liang, Yi Yang, and Alexander G Hauptmann. 2018. RCAA: Relational context-aware agents for person search. In *Proceedings of the European Conference on Computer Vision*. 84–100.

[3] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. 2018. Person search via a mask-guided two-stream cnn model. In *Proceedings of the European Conference on Computer Vision*. 734–750.

[4] Xinlei Chen and Abhinav Gupta. 2017. Spatial memory for context reasoning in object detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 4086–4096.

[5] Zhe Chen, Shaoli Huang, and Dacheng Tao. 2018. Context refinement for object detection. In *Proceedings of the European Conference on Computer Vision*. 71–86.

[6] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2015. Gated feedback recurrent neural networks. In *Proceedings of the International Conference on Machine Learning*. 2067–2075.

[7] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. 2016. R-fcn: Object detection via region-based fully convolutional networks. In *Proceedings of the Advances in Neural Information Processing Systems*. 379–387.

[8] Santosh K Divvala, Derek Hoiem, James H Hays, Alexei A Efros, and Martial Hebert. 2009. An empirical study of context in object detection. In *Proceedings of the IEEE Conference on computer vision and Pattern Recognition*. IEEE, 1271–1278.

[9] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. 2014. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 8 (2014), 1532–1545.

[10] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. 2019. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 6569–6578.

[11] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. 2009. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 9 (2009), 1627–1645.

[12] Carolina Galleguillos, Andrew Rabinovich, and Serge Belongie. 2008. Object categorization using co-occurrence, location and appearance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*. 2961–2969.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[15] Lingxiao He, Yinggang Wang, Wu Liu, He Zhao, Zhenan Sun, and Jiashi Feng. 2019. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. 8450–8459.

[16] Zhenwei He and Lei Zhang. 2018. End-to-end detection and re-identification integrated net for person search. In *Proceedings of the Asian Conference on Computer Vision*. Springer, 349–364.

[17] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7132–7141.

[18] Yukun Huang, Zheng-Jun Zha, Xueyang Fu, Richang Hong, and Liang Li. 2020. Real-world Person Re-Identification via Degradation Invariance Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 14084–14094.

[19] Yangbangyan Jiang, Qianqian Xu, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. 2019. Dm2c: Deep mixed-modal clustering. In *Advances in Neural Information Processing Systems*. 5888–5892.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Proceedings of the Advances in Neural Information Processing Systems*. 1097–1105.

[21] Xu Lan, Xiatian Zhu, and Shaogang Gong. 2018. Person search by multi-scale matching. In *Proceedings of the European Conference on Computer Vision*. 536–552.

[22] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. 2019. Global-local temporal representations for video person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. 3958–3967.

[23] Jianing Li, Shiliang Zhang, and Tiejun Huang. 2019. Multi-scale 3d convolution network for video based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 8618–8625.

[24] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2197–2206.

[25] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 2117–2125.

[26] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2980–2988.

[27] Giuseppe Lisanti, Niki Martinel, Alberto Del Bimbo, and Gian Luca Foresti. 2017. Group re-identification via unsupervised transfer of sparse features encoding. In *Proceedings of the IEEE International Conference on Computer Vision*. 2449–2458.

[28] Hao Liu, Jiashi Feng, Zequn Jie, Karlekar Jayashree, Bo Zhao, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. 2017. Neural person search machines. In *Proceedings of the IEEE International Conference on Computer Vision*. 493–501.

[29] Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, and Meng Wang. 2019. Adaptive transfer network for cross-domain person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7202–7211.

[30] Jiawei Liu, Zheng-Jun Zha, Xuejin Chen, Zilei Wang, and Yongdong Zhang. 2019. Dense 3D-convolutional neural network for person re-identification in videos. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 1s (2019), 1–19.

[31] Jiawei Liu, Zheng-Jun Zha, Richang Hong, Meng Wang, and Yongdong Zhang. 2019. Deep Adversarial Graph Attention Convolution Network for Text-Based Person Search. In *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 665–673.

[32] Jiawei Liu, Zheng-Jun Zha, QI Tian, Dong Liu, Ting Yao, Qiang Ling, and Tao Mei. 2016. Multi-scale triplet cnn for person re-identification. In *Proceedings of the 24th ACM international conference on Multimedia*. 192–196.

[33] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. 2018. Structure inference net: Object detection using scene-level context and instance-level relationships. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6985–6994.

[34] Riccardo Mazzon, Fabio Poiesi, and Andrea Cavallaro. 2013. Detection and tracking of groups in crowd. In *Proceedings of the 10th IEEE International Conference on Advanced Video and Signal Based Surveillance*. IEEE, 202–207.

[35] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller. 2016. Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1325–1334.

[36] Bharti Munjal, Sikandar Amin, Federico Tombari, and Fabio Galasso. 2019. Query-Guided End-To-End Person Search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 811–820.

[37] Woonhyun Nam, Piotr Dollár, and Joon Hee Han. 2014. Local decorrelation for improved pedestrian detection. In *Proceedings of the Advances in Neural Information Processing Systems*. 424–432.

[38] Alexander Neubeck and Luc Van Gool. 2006. Efficient non-maximum suppression. In *Proceedings of the 18th International Conference on Pattern Recognition*, Vol. 3. IEEE, 850–855.

[39] Filip Radenović, Giorgos Tolias, and Ondřej Chum. 2018. Fine-tuning CNN image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 7 (2018), 1655–1668.

[40] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 779–788.

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the Advances in Neural Information Processing Systems*. 91–99.

[42] Weijian Ruan, Wu Liu, Qian Bao, Jun Chen, Yuhao Cheng, and Tao Mei. 2019. Poinet: pose-guided ovonic insight network for multi-person pose tracking. In *Proceedings of the 27th ACM International Conference on Multimedia*. 284–292.

[43] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. 2018. Person re-identification with deep similarity-guided graph neural network. In *Proceedings of the European Conference on Computer Vision*. 486–504.

[44] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. 2019. AANet: Attribute Attention Network for Person Re-Identifications. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7134–7143.

[45] Norimichi Ukita, Yusuke Moriguchi, and Norihiro Hagita. 2016. People re-identification across non-overlapping cameras using group features. *Computer Vision and Image Understanding* 144 (2016), 228–236.

[46] Jimin Xiao, Yanchun Xie, Tammam Tillo, Kaizhu Huang, Yunchao Wei, and Jiashi Feng. 2019. IAN: the individual aggregation network for person search. *Pattern Recognition* 87 (2019), 332–340.

[47] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. 2016. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1249–1258.

[48] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. 2017. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3415–3424.

[49] Yuanlu Xu, Bingpeng Ma, Rui Huang, and Liang Lin. 2014. Person Search in a Scene by Jointly Modeling People Commonness and Person Uniqueness. In

Proceedings of the 22Nd ACM International Conference on Multimedia. 937–940.

[50] Yichao Yan, Qiang Zhang, Bingbing Ni, Wendong Zhang, Minghao Xu, and Xiaokang Yang. 2019. Learning Context Graph for Person Search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2158–2167.

[51] Zheng-Jun Zha, Jiawei Liu, Di Chen, and Feng Wu. 2020. Adversarial attribute-text embedding for person search with natural language query. *IEEE Transactions on Multimedia* 22, 7 (2020), 1836–1846.

[52] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. 2019. Densely Semantically Aligned Person Re-Identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 667–676.

[53] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. 2013. Unsupervised salience learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3586–3593.

[54] Yiru Zhao, Xu Shen, Zhongming Jin, Hongtao Lu, and Xian-sheng Hua. 2019. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4913–4922.

[55] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*. 1116–1124.

[56] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. 2017. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1367–1376.

[57] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. 2015. Towards open-world person re-identification by one-shot group-based verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 3 (2015), 591–606.