Improved Instance Discrimination and Feature Compactness for End-to-End Person Search

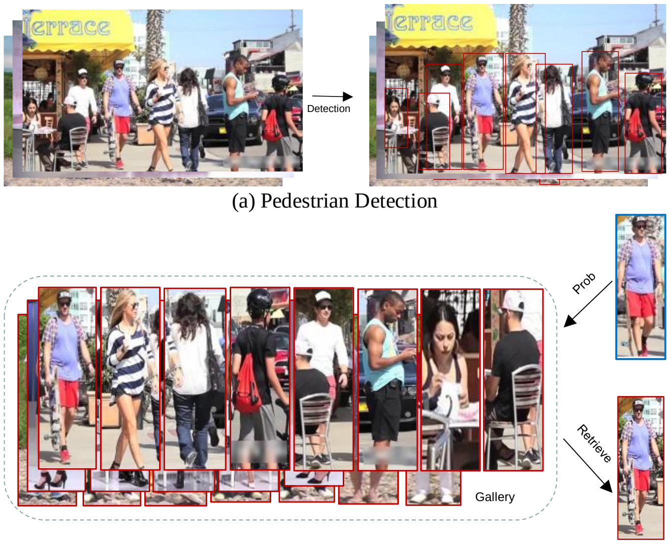
Shaowei Hou*, Cairong Zhao*◆, Zhicheng Chen, Jun Wu, Zhihua Wei, Duoqian Miao

***Abstract*—Person search aims to locate and retrieve specific pedestrians in scene images, including two subtasks, pedestrian detection and person re-identification. Recently, triplet loss has been widely used in person re-identification, which effectively improves the pedestrian features embedding and achieves superior performance. However, forming triplet in the person search is not an easy task. Most of the existing end-to-end person search methods are based on Faster R-CNN. The training process of person re-identification part is affected by the detector. It is difficult to form pedestrian triplets within a limited batch size. Also, there are many pedestrian identities in the person search dataset, but each pedestrian identity only has a few samples. It is difficult to learn a robust pedestrian feature representation for person search. To resolve the problem discussed above, a novel Feature Compactness (FC) Loss for the person search is designed, which efficiently improves the inter-class discrimination and intra-class compactness of pedestrian features embedding without the need for positive or negative pairs. Besides, we propose a pedestrian attention module (PAM) to help the network focuses more on pedestrian information and suppresses irrelevant background information. Our method achieves comparable performance on two benchmarks, CUHK-SYSU and PRW, and achieves 91.96% of mAP and 93.34% of rank1 accuracy on CUHK-SYSU.**

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

I. Introduction

Image retrieval is a fundamental task in the field of computer vision. It has gained much attention in recent years, along with the revival of deep learning. Person search[1] and person re-identification[39,40,41] are image-based retrieval tasks that aim to locate a given person from a bunch of images. Different from person re-identification, person search is an extension of the person re-identification task. Person search locates person from scene images instead of the cropped images. It includes two subtasks: pedestrian detection and person re-identification. The first part/task of the person search task is to locate the pedestrian's position from the scene image. The second part/task is to match and retrieve the target picture based on the positioned pedestrian. The process is shown in Figure 1. Compared with the person re-identification[30,44,45,46] task, the person search task is closer to the real application. However, variation of human pose, camera viewpoint, illumination, occlusion, low resolution, background clutter make the task challenging. Person search has great value in mass video surveillance, including finding missing children, searching criminals, cross-camera pedestrians tracking, etc.



(a) Pedestrian Detection

(b) Person Re-identification

Fig. 1. Comparison of pedestrian detection and person re-identification. In (a) The purpose of pedestrian detection is to locate all pedestrians in the scene image. In (b) The purpose of person re-identification is to retrieve all pictures of a given pedestrian

The existing person search works can be divided into two categories: two-step method[5,8,13,23,52] and end-to-end method[2,3,7,9,11,14]. Two-step methods use a separate pedestrian detection model and person re-identification model to accomplish the person search task. The end-to-end person search method[16,18,19,20,22] combines pedestrian detection and person re-identification into a unified model for training and inference. Most existing two-step[5] work train separate pedestrian detection and person re-identification models. However, such two-step methods are prone to ignore part of the

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62076184, 61673299, 61976160 and 62076182. This work was also supported by Shanghai Innovation Action Project of Science and Technology under Grant 20511100700. This work was also supported by the Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science（East China Normal University), Ministry of Education. This work is also supported by the Fundamental Research Funds for the Central Universities.

* The authors contribute equally to this work. ◆Corresponding author.

Shaowei Hou, Cairong Zhao, Zhicheng Chen, Zhihua Wei and Duoqian Miao are with the Department of Computer Science and Technology, Tongji University, Shanghai, China, E-mail: zhaocairong@tongji.edu.cn .

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

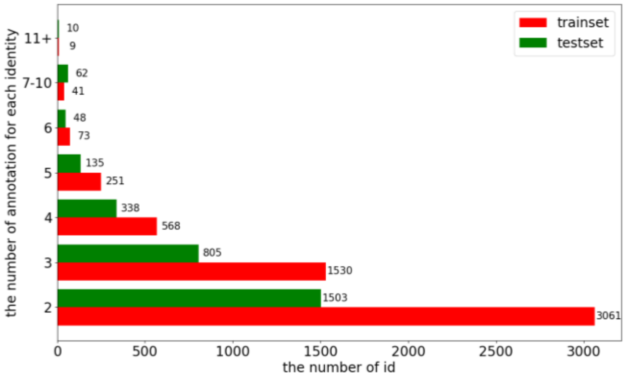


Fig. 2. The number of annotations for each identity on CUHK-SYSU, the red strips indicate trainset data, green strips indicate testset data.

context information, which affects the recognition accuracy of person re-identification. The end-to-end[19] person search method can effectively reduce the model's training time and reduce the interference caused by the object detection model's misdetection.

Person search consists of both pedestrian localization and retrieval tasks that require precise location information and fine-grained appearance information to match. For getting robust pedestrian feature representation, enormous image samples of the same identity are required. However, collecting such kind of training data is not plausible. Moreover, the scale of pedestrians may varies in different camera viewpoints. Facing dramatic scale variance and limited samples for each identity, the person search model needs to learn more discriminative features to distinguish different pedestrians. Conventional softmax loss has been widely used in the field of classification and has achieved certain progress. However, the softmax loss only considers whether the feature embedding classification can be correctly done, while ignore inter-class diversity and intra-class similarity. As a result, using softmax loss in person search would lead to performance deterioration under drastic intra-class appearance variations situation. Existing end-to-end person search efforts to improve the feature distribution through the introduction of center loss[10], which makes intra-class features more compact but ignores the discreteness of inter-class features. Recently, triplet loss[42,43] has been widely used in person Reid, which can achieve superior performance. However, for the end-to-end person search methods implemented based on Faster R-CNN[32], forming triplet within the input limit batch in the person search is not an easy task [10]. As shown in Figure 2, there are many pedestrian identities in the person search dataset, while each pedestrian identity only has a few samples. There are random, unduplicated and uneven for the pedestrians appearing in each frame and it is hard to form a balanced number of positive pedestrian pairs as negative pairs in a limited batch of Faster R-CNN inputs[10]. Therefore, end-to-end person search cannot use triplet loss to improved embedding feature distributions.

To address the above issues, a novel Feature Compactness (FC) Loss for the person re-identification subtask in the person search. It is designed to enhance pedestrian instance discrimination and improve the pedestrian features' compactness in multi-scale scenes. Specifically, the purpose of FC loss is to make intra-class feature compact, which pushes the sample to the corresponding positive center, followed by the inter-class discrepancy, which pushes the sample away from all other negative centers. Notably, employing FC loss can avoid the need to aggregate positive and negative pairs; Besides, we design a Pedestrian Attention Module (PAM) to leverage the spatial relationship between features to generate a spatial attention map, effectively learning the person's feature representation and suppressing the distraction from irrelevant information. The proposed network for Improved Instance Discrimination and Feature Compactness(IIDFC) is jointly optimized by PAM and FC loss for the model, which learns more discriminative feature representations and improves the distribution of pedestrian features, resulting in more compact intra-class and more discriminative inter-class distribution. Comprehensive experiments show that our proposed method has achieved competitive performance on two person search datasets, CUHK-SYSU and PRW.

In summary, the contribution of our work is threefold.

1) A Feature Compactness (FC) loss is designed to enhance the inter-class discriminability and intra-class compactness of pedestrian features in Reid's head, which avoids selecting positive and negative sample pairs.
2) A pedestrian attention module (PAM) with shared weights for multiple receptive fields is designed to discover spatial relationships and generate pedestrian correlated features and suppress irrelevant information.
3) We demonstrate the efficiency of our method with comprehensive experiments and achieved competitive performance on the CUHK-SYSU and PRW datasets.

II. Related work

Person Search. The task of person search is to find the query person from natural scene images. It includes searching (pedestrian detection) and matching (person re-identification). Existing works can be mainly divided into two categories: end-to-end methods and two-step methods. Two-step methods train isolated models for pedestrian detection and person re-identification, while end-to-end fuse the two tasks into a single model.

For two-step approaches, Zheng et al.[4] proposed a new person dataset and proposed to reweight the similarity of Reid with the detection confidence. Chen et al.[8] introduced mask information to augment the pedestrian features and proposed to use the background information as the auxiliary information for pedestrian matching. Han et al.[13] proposed a differentiable ROI that combined the detection task with the Reid task and trained the model with the proxy triplet loss. Considering the objectness and repulsion information, Yao et al.[52] proposed OR similarity. On the contrary, Xu et al. [1] believe that sequential combination is not an optimal solution for the person search task and propose an end-to-end framework to optimize pedestrian detection and person re-identification in a unified

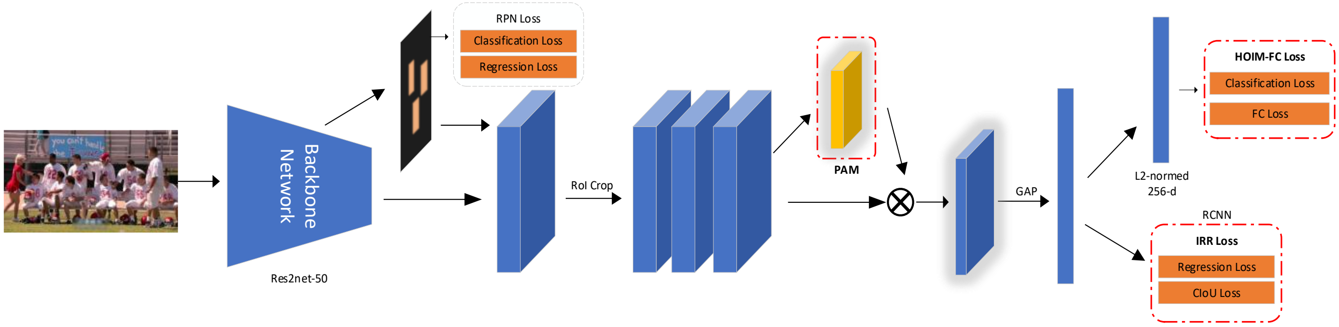


Fig. 3. Overview of our framework. Our network is based on Faster R-CNN with a Res2Net50. The framework includes three modules: feature extraction network, pedestrian attention module, and Reid network. We project the pedestrian depth embedding features into an L2-normalized 256-d subspace, and then train it with a proposed Feature Compactness loss. PAM denotes Pedestrian Attention Module.

framework.

End-to-end person search approaches are mainly based on Faster R-CNN [32]. Early works simply connect an auxiliary linear layer to the top convolutional layer of Faster R-CNN to generate Reid embeddings. The whole model was jointly trained under the supervision of standard Faster R-CNN loss, and OIM [2] loss or Center loss [10]; Liu et al. [3] proposed recursively narrowing the search region, where detectors can directly match on panoramic images instead of generating a large number of frames on the image. Yan et al. [7] also adopted a similar idea and constructed a graph learning network using contextual information. Chang et al. [11] reduce redundant proposals and computational overhead by training relational context-aware agents. Munjal et al. [9] merged query information into a conjoined network to guide feature learning, suggestion generation, and similarity calculation. Although the above end-to-end methods achieve considerable performance, they do not focus enough on pedestrian feature distribution. Xiao et al. [10] proposed to jointly supervise the model's training with center loss and softmax loss, which improves the compactness of intra-class features to some extent but does not consider the discreteness of inter-class. Since end-to-end person search is affected by the batch size and data distribution to find suitable positive class sample pairs, the traditional triplet loss cannot be directly applied to the end-to-end person search task. To solve the problems above, we propose FC loss, which can avoid selecting positive sample pairs, make the samples of intra-class closer to each other, and have a certain inter-class margin.

Pedestrian Detection. Pedestrian detection is a sub-field of object detection whose key object detection objective is to locate targeted objected in scene images. Pedestrian detection is a classic application in object detection. Autonomous driving, video surveillance, criminal investigation, and other fields have received extensive attention. Early pedestrian detection methods using detectors such as HOG[33] and ICF[34] have achieved definite progress efficiently. In recent years, with the development of deep learning in detection, some generic object detection approaches, such as Faster RCNN, have been introduced into pedestrian detection, which has extensively boosted pedestrian detection development. However, the low resolution of the features after convolution leads to insufficient detection of pedestrians on small-scale pedestrians. Many recent works have proposed methods such as feature fusion, high-resolution hand-crafted features, and multi-resolution comprehensive detection results to address this problem. Some pedestrian detection works focus on pedestrian detection in crowded scenes. Wang et al. [24] proposed Repulsion loss, which improves pedestrian detection accuracy in dense scenes and makes bounding box regression more accurate. Pang et al. [31] proposed mask guided attention method to solve the problem of low pedestrian detection effect partially due to occlusion. Lin et al. [48] proposed a graininess-aware deep feature learning method for pedestrian detection, which incorporates fine-grained information into convolutional features for better discrimination of body parts. These different tasks have promoted the development of the field of pedestrian detection to varying degrees.

Person Re-identification. The task of person re-identification [26,27,28,29,54,55,59] is to retrieve specific pedestrians from the cropped gallery images. Traditional person re-identification methods focus on measuring manually designed features. There is some work dedicated to metric learning, such as Hu et al.[47] proposed a sharable and individual MvML approach. With the development of deep learning in recent years, many person re-identifications[50, 56,57,58] based on deep learning have made significant progress and can be roughly divided into two categories. One is to propose different model structures for extracting pedestrians with more fine-grained details features, such as PCB[15] model and MGN[17] model; the other is to learn better fine-grained classification features by proposing different loss functions[42,43], such as center loss, triple loss, etc. Liu et al.[51] propose PrGCN) method to improve the similarity measurement. These losses pulling in the distance between similar samples and pushing away the distance between dissimilar samples, i.e., increasing the interclass distance and decreasing the intraclass distance. Similarly, the hybrid multinomial FC loss proposed in this paper also learns a feature representation that is more conducive to inter-class sparsity and intra-class compactness.

III. Proposed Method

In this section, we introduce the proposed method in detail. The overall framework is elaborate in subsection 3.1; The proposed pedestrian attention module is introduced in

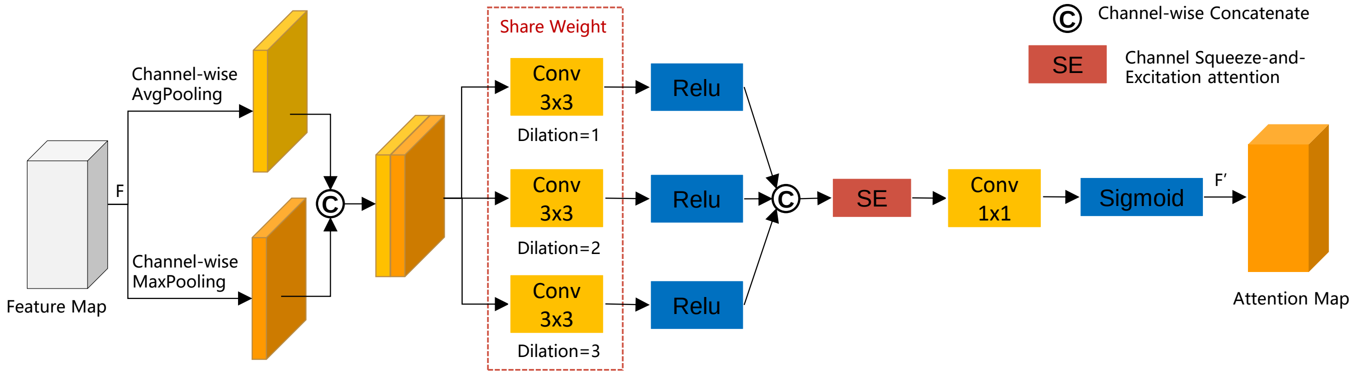


Fig. 4. The network architecture of our Pedestrian Attention Module(PAM). The general process of PAM, firstly the feature map is subjected to average pooling and maximum pooling operations respectively, followed by sampling through three dilation convolutions with shared weights respectively, then the network is allowed to learn the effect of different convolutions adaptively through SE, followed by outputting to an attention map.

subsection 3.2; Finally, a Feature Compactness loss is introduced in subsection 3.3.

3.1 Model Overview

As shown in Figure 3, the overall framework of our end-to-end model is based on Faster RCNN, which is mainly composed of three parts: a backbone network, a regional proposal network, and a Reid network. The backbone network is composed of the top four modules of res2net and is used for feature extraction; The regional proposal network is used to generate pedestrian candidate proposal; The Reid network is used for pedestrian identification. The Reid network contains the last module of res2net and a pedestrian attention module, which is used to extract discriminative person Reid features. To learn more discriminative pedestrian feature representations, firstly, the pedestrian attention module is designed, which sampled pedestrian features with three shared weights convolution layer with different dilation rates, and then the Squeeze-and-Excitation(SE) module makes the network adaptively select optimal receptive fields; Hence more pedestrian fine-grained feature representations can be learned and the learning of irrelevant information is suppressed; Secondly, the spatial distribution of pedestrian features is improved by designing FC loss, which makes intra-class features more compact and inter-class features more discriminative.

In the training stage, the backbone network extracts the image level feature from the input image. On top of the image-level features, the RPN network generates pedestrian proposals. The proposal regions are then pooled by RoI pooling and the pooled features are sent to the Reid network for identification. In the Reid network, the input features are normalized to unit length and optimized with FC loss. To make the bounding box regression more accurate, the CIoU[35] loss and the original SmoothL1 loss were introduced as the regression supervision for the RCNN part. The Improved RCNN Regression (IRR) loss of the RCNN head is shown as follows,

$$\mathcal{L}_{irr\_reg} = \gamma\mathcal{L}_{smooth_{l1}} + \beta\mathcal{L}_{CIoU}, \qquad (1)$$

where CIoU loss not only handles the overlap area and central point distance of bounding boxes at the same time but also consider aspect ratios for bounding boxes, the γ and β are tun-able hyperparameters,

$$\mathcal{L}_{CIoU} = 1 - \mathrm{IoU} + \frac{\rho^2(b,b^{gt})}{c^2} + \alpha v, \qquad (2)$$

$$\alpha = \frac{v}{(1-IoU)+v}, \qquad (3)$$

$$\nu = \frac{4}{\pi^2}\left(arctan\frac{\omega^{gt}}{h^{gt}} - arctan\frac{\omega}{h}\right)^2. \qquad (4)$$

where α is a positive trade-off coefficient and ν measures the consistency of aspect ratio, where b and b^{gt} denote the central points of the predicted box and the ground-truth box, $\rho(\cdot)$ is the Euclidean distance, and c is the diagonal length of the smallest enclosing box covering the two boxes, the w and h denote the width and height of the prediction boxes, respectively.

With the optimization of $\mathcal{L}_{irr\_reg}$ loss, the upstream pedestrian detection task can produce more accurate bounding boxes, and the downstream Reid model can generate more discriminative feature embeddings due to the refined pedestrian location information obtained.

Our model generates multiple pedestrian bounding boxes and corresponding Reid features from the input scene images during inference. Then, we rank the person from gallery images according to their cosine distances to the target person.

3.2 Pedestrian attention module.

In real scene images, target pedestrians are often occluded by irrelevant people and objects, thus causing local occlusion. Many models[6] hope to find out more discriminative color or texture related features in the image through attention mechanism and ignore the irrelevant background features. Therefore, we propose a parameter-economic and efficient module that a joint spatial and channel pedestrian attention module (PAM). PAM mainly focuses on the location of the significant feature, which enhances the information of pedestrians and suppresses the learning of irrelevant information, making the model focus more on informative parts of a pedestrian. The pedestrian attention module is shown in Figure 4.

Earlier experiments[6] have shown that applying average-pooling and max-pooling operations on the channel axis can effectively highlight the foreground region's information. Similarly, in the PAM module, we first perform average pooling and maximum pooling operations on the channel axes and concatenate them to generate an efficient feature descriptor. We apply three dilation convolution layers of shared weights with ReLU as activation function to the

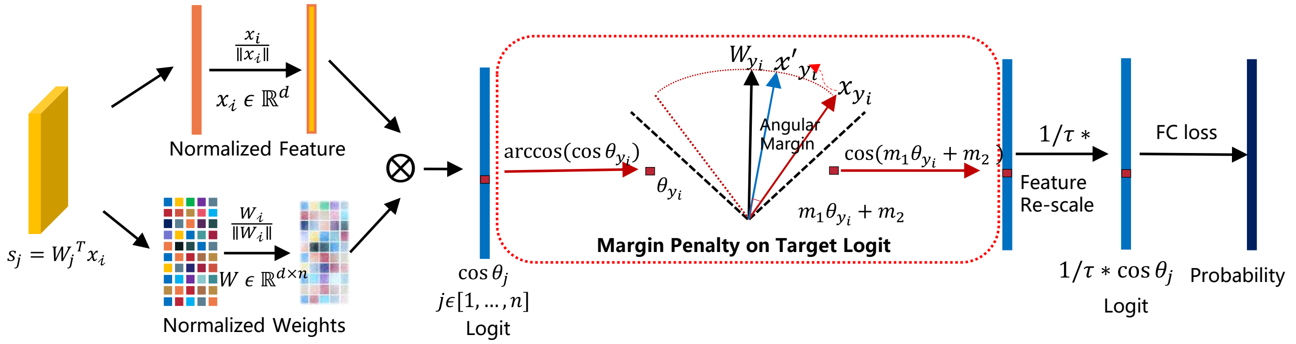


Fig. 5. Overview of FC loss. We first normalized the feature x_i and weight W by L_2normalization, then we get the $\cos\theta_j$ (logit) for each class as $W_j^T x_i$. By the inverse cosine operation $\arccos(\cos\theta_{y_i})$ on $\cos\theta_{y_i}$, we get the angle between the embedded features x_i and the ground truth weights W_{y_i}. Where the W_{y_i}denote a kind of center for each class. Secondly, we add a penalty factor m_1 and an angular margin m_2 to target (ground truth) angle θ_{y_i} to obtain $m_1\theta_{y_i} + m_2$. Finally, the final logit is obtained by calculating the cosine function $\cos(m_1\theta_{y_i} + m_2)$ and rescaled by the temperature sensitivity factor τ.

concatenated feature descriptor. Then , channel-wise Squeeze-and-Excitation (SE) [38] is utilized to learn the significance of feature maps from different dilation rates adaptively. Finally, PAM generate a pedestrian attention map $M_p(\mathrm{F}) \in R^{H \times W}$. It is worth noting that the three parallel dilation convolutions share weights and the dilation rate of the three convolutions are 1,2,3, respectively. The dilation convolution enlarges the size of the convolution kernel, i.e., increases the receptive field, but there is no increase in the number of parameters and computations by sharing the weights. Specifically, the receptive field of a dilated 3×3 convolution can be calculated by $3 + 2(d_s - 1)$, where d_s is the dilation rate. The detailed operation of PAM can be expressed as:

$$F' = M_p(\mathrm{F}) \otimes \mathrm{F}, \tag{5}$$

$$M_p(\mathrm{F}) = \delta(f^{1\times1}(\mathrm{S}(\,[C_1; C_2; C_3]))), \tag{6}$$

$$C_i = \sigma(f^{3\times3}_{d_s=i}([F^p_{avg}; F^p_{max}])), \tag{7}$$

where $M_p(\mathrm{F})$ denotes the PAM. The δ denotes the sigmoid function and σrepresents the ReLU activation function, the $f^{1\times1}$ means a convolution operation with a kernel size of 1×1. The $\mathrm{S}(\cdot)$ denotes channel attention SE operation, F^p_{avg} and F^p_{max} denote the feature maps after average pooling and maximum pooling, respectively, C_i indicates the feature map sampled by convolution with different dilation rates d_s and fixed filter size of 3×3 and activated by the Relu function.

3.3 Feature Compactness loss

There are many pedestrians with unknown identities in the person search task, and the number of samples for each identity is limited. It is difficult for the traditional softmax to exploit these using this unlabeled identity information, as they have no specific class-id. With the unlabeled data to learn better features representation, many person search tasks are based on OIM[2] loss,

$$\mathcal{L}_{OIM} = -\frac{1}{T}\sum_{i=1}^{T} \log \frac{e^{s_i/\tau}}{\sum_{j=1}^{N} e^{s_j/\tau} + \sum_{k=1}^{M} e^{s_k/\tau}} \tag{8}$$

where $\mathrm{s} = W^T x_i \in \mathbb{R}^{N+M}$. In contrast to the projection matrix of the softmax layer, the W is parameter an external buffer, and it does not affect the update of the model parameters. The OIM loss maintains an online lookup table (LUT) $\mathrm{V} \in \mathbb{R}^{D \times N}$ and a circular queue (CQ) $\mathrm{U} \in \mathbb{R}^{D \times M}$, where D, N denote the feature dimension and the number of pedestrian identities in the train set, respectively. M is a fixed value close to N. The LUT and external buffer CQ are used to store the features of all labeled identities and unlabeled identities, respectively.

The HOIM[20] loss further enhances the OIM loss by highlighting the distinction between foreground and background. A background feature buffer is added to the OIM loss as a bounding box classification in the Faster R-CNN. HOIM also improved the buffer update strategy by Selective Memory Refreshment[20], the SMR strategy also enhances HOIM's discriminatory power. To further aggregate the pedestrian detection and person reidentification tasks in HOIM, the first layer mainly describes the goal of pedestrian detection, which addresses the distinction between foreground and background; the second layer describes the task of person reidentification. HOIM loss can be defined as,

$$p_{f\_i} = \sum_{i=1}^{N+M} \frac{e^{s_i/\tau}}{\sum_{j=1}^{N+M+B} e^{s_j/\tau}} \tag{9}$$

$$p_{b\_i} = \sum_{i=N+M+1}^{N+M+B} \frac{e^{s_i/\tau}}{\sum_{j=1}^{N+M+B} e^{s_j/\tau}} \tag{10}$$

$$\mathcal{L}_{rcnn\_cls} = -y\, log\big(p_{f\_i}\big) - (1-y)\, log\big(p_{b\_i}\big) \tag{11}$$

$$\mathcal{L}_{HOIM} = \mathcal{L}_{rcnn\_cls} + \lambda \mathcal{L}_{OIM} \tag{12}$$

where the p_{f_i} and p_{b_i} denote x predict foreground probability and background probability, respectively. The $\mathcal{L}_{rcnn\_cls}$ is the first level of HOIM loss, which can be formulated as a binary cross-entropy loss. $\lambda = 2(\sum_{i=1}^{N+M} \frac{e^{s_i/\tau}}{\sum_{j=1}^{N+M+B} e^{s_j/\tau}})^2$ is the loss weight for $\mathcal{L}_{OIM}$. It changes according to the detection confidence digit. y, which is a binary label indicating whether x is a person or not.

Our FC loss takes a further step based on HOIM loss by considering inter-class discrimination and intra-class compactness. As illustrated in Table I and Figure 5. First, We transform the logit as $\mathrm{s} = W_j^T x_i = \| W_j \| \| x_i \| \cos\theta_j$, where θ_j is the angle between the weight W_j and the depth embedding feature x_i. Then, the depth pedestrian feature and the weight parameters of the class center are L_2 normalized and dotted product to obtain $\cos\theta_j$. It is worth noting that the normalization of features and weights makes the prediction depend only on the angle θ_j. The angle $\theta_j$is obtained by the inverse cosine function, and the penalty factor m_1 and the angle margin m_2 are added to θ_j, i.e., $m_1\theta_{y_i} + m_2$. After that, we obtain the target logit again by the cosine function, and all the logits are re-scaled by a fixed temperature sensitivity factor τ. The FC loss can be defined as,

$$\mathcal{L}_{FC} = -\frac{1}{T}\sum_{i=1}^{T} \log \frac{e^{(\cos(m_1\theta_{y_i}+m_2))/\tau}}{e^{(\cos(m_1\theta_{y_i}+m_2))/\tau}+\sum_{j=1,j\neq y_i}^{N+M} e^{\cos\theta_j/\tau}} \quad (13)$$

where θ_{y_i} denotes the angle between x_i and W_{y_i} while θ_j denotes the angle between x_i and W_j . m_1 and m_2 is a hyper-parameter indicating the FC loss and τ is a temperature factor in transforming the softness of the probability distribution. Table I illustrating the decision boundary in the binary classification case for several common losses.

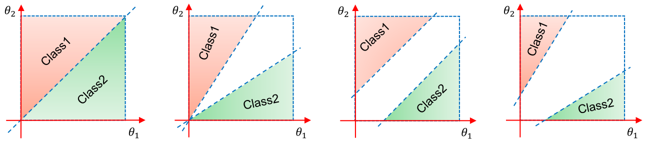
$$\mathcal{L}_{HOIM-FC} = \mathcal{L}_{rcnn\_cls} + \lambda\mathcal{L}_{FC} \quad (14)$$

where $\mathcal{L}_{rcnn\_cls}$ is the first level of HOIM-FC loss and the $\lambda = 2(\sum_{i=1}^{N+M}\frac{e^{s_i/\tau}}{\sum_{j=1}^{N+M+B} e^{s_j/\tau}})^2$ is the loss weight for $\mathcal{L}_{FC}$.

We visualize the decision boundaries of FC loss and several commonly used loss functions under the binary classification setting. The decision boundary in softmax loss can be defined as $(W_1 - W_2)\ x = 0$, where W_j represents weights of the j-th class in softmax loss. We transform the logit as $s = W_j^T x_i = \parallel W_j \parallel \parallel x_i \parallel \cos\theta_j$, where θ_j is the angle between the weight W_j and the feature x_i. We fix the individual weights W_j and embedding feature x_i by L_2 normalization to $\parallel W_j \parallel = 1$ and $\parallel x_i \parallel = 1$, respectively. So that the decision boundary of softmax loss becomes $\cos(\theta_1) - \cos(\theta_2) = 0$. In the binary class case, by transformation, the new decision boundary only depends on θ_1 and θ_2 and the decision boundaries of FC loss for class 1 and class 2 become $\cos(m_1\theta_1 + m_2) - \cos(\theta_2) = 0$ and $\cos(\theta_1) - \cos(m_1\theta_2 + m_2) = 0$, respectively. Both the OIM loss and the HOIM loss apply softmax cross entropy to the logits predicted, so the decision boundary of OIM loss and HOIM loss in the binary classification case the same as softmax loss.

TABLE I
COMPARISON OF DIFFERENT LOSSES DECISION BOUNDARIES IN BINARY CASE.

Loss Function	*Decision Boundary*
Softmax Loss	$cos\ \theta_1 - cos\ \theta_2 = 0$
L-Softmax Loss[49]	$\parallel x \parallel (cos\ m\theta_1 - cos\ \theta_2) = 0$ *for class 1* $\parallel x \parallel (cos\ \theta_1 - cos\ m\theta_2) = 0$ *for class 2*
SphereFace Loss[25]	$cos\ m\theta_1 - cos\ \theta_2 = 0$ *for class 1* $cos\ \theta_1 - cos\ m\theta_2 = 0$ *for class 2*
Arcface Loss[37]	$cos(\theta_1 + m) - cos\ \theta_2 = 0$ *for class 1* $cos\ \theta_1 - cos(\theta_2 + m) = 0$ *for class 2*
FC Loss	$cos(m_1\theta_1 + m_2) - cos(\theta_2) = 0$ *for class 1* $cos(\theta_1) - cos(m_1\theta_2 + m_2) = 0$ *for class 2*



(a) Softmax (b) SphereFace (c) ArcFace (d) FC(ours)

Fig. 6. The decision boundaries of several generic loss functions such as Softmax, SphereFace, ArcFace and FC loss in the binary classification case are illustrated, where the dashed lines indicate the decision boundaries and the white areas represent the decision margins

To further illustrate our FC loss's superiority, we quantify the feature space of the binary classification, assuming that the feature space of the binary classification is on the contour of a circle with $redius = 1$. Then the distance to the two-class centers can be expressed as θ_1 and θ_2, where $\theta_1, \theta_2 \in (0, \pi)$. Similarly, the decision boundary of several other common losses can be calculated, and the visualization results are shown in Figure 6. The decision bounds for several common losses under binary classification are shown in Table I. In this binary classification example, the FC loss has a smaller decision area compare to other commonly used losses, making the intra-class features more compact. It is also shown that the proposed FC loss can effectively improve the distribution of different classes of features and enhance the discriminative of different instances.

This framework uses Faster RCNN as the detector, and the total loss function of our framework is shown as follows,

$$\mathcal{L}_{all} = \lambda_1\mathcal{L}_{rpn\_cls} + \lambda_2\mathcal{L}_{rpn\_reg} + \lambda_3\mathcal{L}_{irr\_reg} + \lambda_4\mathcal{L}_{HOIM-FC} \quad (15)$$

where $\mathcal{L}_{rpn\_cls}$ and $\mathcal{L}_{rpn\_reg}$ are the classification loss and regression loss of the RPN module in the Faster RCNN detector, $\mathcal{L}_{rpn\_reg}$ is the Smooth $\mathcal{L}_1$ loss, $\mathcal{L}_{irr\_reg}$ are refinement regression loss of the second stage RCNN in the Faster RCNN detector, the $\mathcal{L}_{HOIM-FC}$ loss contains $\mathcal{L}_{rcnn\_cls}$ and $\mathcal{L}_{FC}$ losses.

IV. EXPERIMENTS

In this section, we first give a brief overview of the existing person search datasets and the evaluation metrics. Then, we describe our implementation in detail. Finally, we report our experimental results and compare them with the state-of-the-art methods.

4.1 Datasets and Evaluation Protocol

Our method is evaluated on two representative datasets CUHK-SYSU [2] and PRW [3]. The details are shown in Table II.

CUHK-SYSU. The dataset is collected from urban street cameras and movie scenes. It contains 18184 images, 8432 pedestrians, and 96143 pedestrian annotated frames. The data set is divided as follows: the training set contains 11206 with 5532 pedestrian identities and 55,272 pedestrian annotations; the test set contains 6,978 images in total, with 2900 pedestrians' identities 40,871 pedestrian annotations. The data set has highly variable in illumination, viewpoint, low resolution, occlusion, and background clutter, representing the environmental complexity and variability of real application scenarios.

PRW. The dataset contains 11,816 video data frame images, a total of 43,110 pedestrian annotated boxes, of which 34,304 annotated boxes have 932 pedestrian category labels. The data set is divided into a training set containing 482 pedestrians with 5,134 scene images and a test set containing 2,057 query pedestrian images and 6,112 candidate scene images.

TABLE II
STATISTICS OF THE DATASET, BOUNDING BOXES SCALE AND EVALUATION SETTING OF CUHK-SYSU AND PRW. BOXES MEANS BOUNDING BOX.

Dataset	Bboxes Scales	All			Train			Test		
		Frames	Boxes	Identities	Frames	Boxes	Identities	Frames	Boxes	Identities
CUHK-SYSU	37×13 ~ 793×297	18,184	96,143	8,432	11,206	55,272	5,532	6,978	40,871	2900
PRW	58×21 ~ 777×574	11,816	43,110	932	5,704	18,048	482	6,112	25,062	450

Evaluation Protocol. In the experiments, the performance of the proposed method is evaluated by two metrics: cumulative matching curve (CMC top-K) and mean average precision (mAP). All the experiments are performed in a single query setting.

TABLE III
THE RESULTS OF IMPLEMENT HOIM METHOD AND THE IMPROVED BACKBONE NETWORK ON CUHK-SYSU AND PRW.

Method	CUHK-SYSU		PRW	
	Rank-1	mAP	Rank-1	mAP
HOIM-ours-implements	89.52	88.88	76.71	37.79
HOIM-base	91.59	90.60	78.90	37.88

TABLE IV
COMPARISON WITH ONE STEP(END-TO-END) AND TWO STEP STATE-OF-THE-ART METHODS ON TWO PERSON SEARCH DATASETS. CUHK-SYSU AND PRW.

Method		CUHK-SYSU		PRW	
		Rank-1	mAP	Rank-1	mAP
Two Step	MGTS[8]	83.70	83.00	72.10	32.60
	CLSA[5]	88.50	87.20	65.00	38.70
	RDLR[13]	94.20	93.00	70.20	42.90
	TCTS[23]	95.10	93.90	87.50	46.80
	OR[52]	93.83	93.23	71.51	52.30
One Step(end-to-end)	OIM[2]	78.70	75.50	49.90	21.30
	NPSM[3]	81.20	77.90	53.10	24.20
	RCAA[7]	81.30	79.30	/	/
	QEEPS[9]	89.10	88.90	76.70	37.10
	CGPS[11]	86.50	84.10	73.60	33.40
	APNet[16]	89.30	88.90	81.40	41.90
	BINet[22]	90.70	90.00	81.70	45.30
	IGPN[14]	91.40	90.30	87.00	47.20
	BPNet[18]	90.50	88.40	**87.90**	**48.50**
	NAE+[19]	92.90	**92.10**	81.10	44.00
	HOIM[20]	90.83	89.74	80.36	39.77
	OIM(ours)	85.97	86.19	72.10	32.51
	OIM(ours)+PAM+FC	88.10	87.24	78.71	38.88
	HOIM-base	91.59	90.60	78.90	37.88
	IIDFC(ours)	**93.34**	91.96	83.37	43.43

4.2 Implementation Details

In our framework, our implementation is based on PyTorch. The model is based on a Faster RCNN detector with a Res2Net-50 backbone and initialized with ImageNet pre-trained weights. The model is trained with 4 NVIDIA GeForce RTX 2080Ti GPUs using model parallel. The training batch size is set to 5, the initial learning rate is 0.003, which is gradually warmed-up at the first epoch and decayed by a factor of 0.1 at the 16th epoch, and the model is trained at epoch 23. SGD with momentum is used to optimize the model. The size of the input training image is resized to a maximum of 900 pixels on the short side and 1500 pixels on the long side. τ is set to 1/30 and in equation (11), m_1 and m_2 are set to 0.9 and 0.1, respectively. In equation (1), the γ and β are initialized to 0.3 and 0.7, the τ is set to 1/30 and in equation (13), m_1 and m_2 are set to 0.9 and 0.1, respectively. In equation (15), $\lambda_1$to λ_4 are set to 1, 5, 5, and 1, respectively. The embedding projection layers are followed by batch normalization layers in the RCNN head.

4.3 Comparisons with State-of-Arts

In this session, we compare with the current methods that have achieved better results in person search to prove the effectiveness and robustness of our proposed method and module. We compare our method with the state-of-the-art methods on two representative datasets, including two step methods and end-to-end methods.

Constructing a strong baseline. We constructed a new baseline, as shown in Table III. We replaced the original backbone in the HOIM method. We also selected more appropriate anchor[53] , the anchor ratios is [0.9, 1.4, 1.9, 2.3, 2.5, 2.7, 2.9, 3.1, 3.5, 4.0, 4.6, 5.2] and anchor scales is [32, 64, 128, 256, 512] The new baseline we constructed achieves Rank-1/mAP of 91.59%/90.60% and 78.90%/ 37.88% on CUHK-SYSU and PRW, respectively. Compared with the previous ones, Rank-1 and mAP are improved by +2.07%, +1.72% and +2.19%, +0.09%, respectively.

Evaluation on CUHK-SYSU. Table IV shows the result of our method and existing state-of-the-art works. Our IIDFC achieves 93.34% Rank-1 accuracy and 91.96% mAP, which is superior to state-of-the-art end-to-end methods. Compared to one of the most competitive methods HOIM, IIDFC outperforms it by +2.51% Rank-1 accuracy and +2.22% mAP. We re-implemented OIM method and added our proposed module to the OIM(ours), and the experiments show that our proposed PAM and FC loss have different degrees of improvement on both person search datasets, further verifying the effectiveness of our proposed method, and the specific

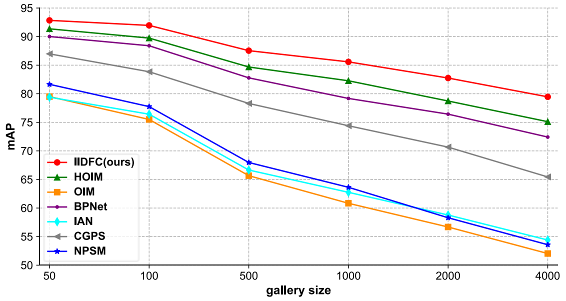


Fig. 7. Performance comparison of mAP on CUHK-SYSU with different gallery sizes of [50,100,500,1000,2000,4000]. The solid red line is our proposed method IIDFC, which performs well under different gallery sizes.

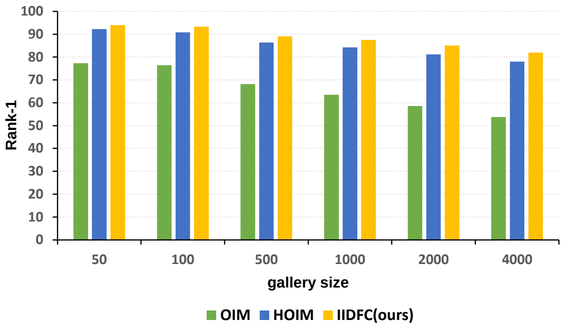


Fig. 8. The Rank-1 comparison on CUHK-SYSU with different gallery sizes of [50,100,500,1000,2000,4000]. There is a considerable gap between the other approaches and our IIDFC.

experimental results can be seen in Table IV.

Evaluation on PRW. Table IV shows the result of our method on PRW. Our model achieves the competitive performance in both mAP and top-1 accuracy among end-to-end methods. Our IIDFC achieves 83.37% Rank-1 accuracy and 43.43% mAP, Compared with HOIM's Rank-1 and mAP 3.01% and +3.66%respectively. Both IGPN and BPNet networks introduce additional information. The IIDFC network is designed with FC loss and PAM attention without introducing additional information, only global information is used for metrics and no additional information is used for alignment, so the performance improvement of the IIDFC network is limited.

Evaluation of different gallery sizes on CUHK-SYSU. We also evaluated the robustness of our method under different gallery sizes on CUHK-SYSU. We compared the mAP of various end-to-end methods under different gallery sizes in Figure 7. IIDFC shows better robustness under different gallery sizes and surpasses the state-of-the-art methods by a large margin. We also evaluate the Rank-1 performance under different gallery sizes and the result is shown in Figure 8. Similarly, the IIDFC has better robustness state-of-the-art methods. Overall, IIDFC generates discriminative feature representations and is more robust under different gallery sizes. Compare to the baseline method HOIM, IIDFC improved the performance of Rank-1 at different gallery sizes, which confirms the effectiveness and robustness of our proposed IIDFC.

Visualization on CUHK-SYSU and PRW. We also visualize the retrieving results and some examples are shown in Figure 10 and Figure 11. Rank-1 person search matches on the CUHK-SYSU and PRW test data are reported. Compared with the baseline method HOIM, it shows that IIDFC leads to the best performance with the correct matching ranked top. Compared with the HOIM method, our proposed method IIDFC can learn more discriminative features for small objects. The visualization results in rows 3,4 of Figure 10 and rows 2,3 of Figure 11 show that IIDFC is more discriminative than the HOIM method by learning some small target pedestrians' features.

TABLE V
PERFORMANCE WITH EACH MODULE OF PROPOSED METHODS ON CUHK-SYSU AND PRW.

Method	CUHK-SYSU		PRW	
	Rank-1	mAP	Rank-1	mAP
HOIM-base	91.59	90.60	78.90	37.88
+IRR	92.31	90.95	79.19	40.22
+ FC	92.45	91.33	81.77	41.11
+PAM	93.34	91.96	83.37	43.43

TABLE VI
COMPARISON OF DIFFERENT BACKBONE

Method	PRW	
	Rank-1	mAP
HOIM-ours-implements(ResNet50)	76.71	37.79
ResNet50+IRR+FC+PAM	81.92	39.81
Res2Net50+IRR+FC+PAM	83.37	43.43
Res2Net50_v1b_26w_4s+IRR+FC+PAM	82.89	43.10

4.4 Ablation Study

We conduct extensive ablation studies to analyze each component of the proposed IIDFC. We evaluate the impact of different modules, ablation studies are conducted on CUHK-SYSU and PRW.

Ablation experiment of different modules. The ablation experiment is based on the new baseline HOIM-base and the test results are shown in Table V. The improved RCNN regression (IRR) loss improves Rank-1/mAP by +0.72%/+0.35% and +0.29%/+2.34% on CUHK-SYSU and PRW, respectively. On this basis, we added the FC loss module, and its gain on the two data sets were (+0.14%, +0.38%) and (+2.58%, +0.89%) of Rank-1 and mAP, respectively. Finally, we added our PAM. The location information of pedestrians is effectively noticed, thereby effectively improving the performance of recognition. The gains on the two data sets are (+0.89%, +0.63%) and (+1.6%, +2.32%) of Rank-1 and mAP, respectively. Based on the baseline, our proposed framework's performance gains on the two data sets are (+1.75%, +1.36%) and (+4.47%, +5.55%) respectively. Our proposed method can effectively improve the discrimination between different instances and enhance the compactness of the same category, to better distinguish between different pedestrian instances.

Comparison of different backbones. To evaluate the generality of our method, we further conduct experiments with

TABLE VII
COMPARISON OF DIFFERENT MODULE RESULTS ON PRW.

Modules	PRW	
	Rank-1	mAP
HOIM-base+IRR+ FC (baseline)	81.77	41.11
baseline+FC+SA	82.47	42.47
baseline+FC+CBAM	81.72	42.00
baseline+FC+Non-local	82.89	42.75
baseline+FC+PAM(ours)	83.37	43.43
baseline+PAM+SphereFace loss	82.30	40.77
baseline+PAM+Arcface loss	82.79	42.80

different backbones on the PRW dataset and the results are shown in Table VI. The three rows show the result of our reimplementation of HOIM, the proposed method on ResNet50, Res2Net50[36], a Res2Net variant, respectively. The results show that the proposed method can bring consistent improvement on the different backbone, which shows the generality of the proposed method.

Comparison of different attention modules. We conducted a set of comparative experiments to validate the effectiveness of our proposed pedestrian attention module. The Experiment results are shown in the first five-row of Table VII. The first row shows the result of our method without any attention module, which is used as the baseline for comparison. The second row to the fifth row shows the result of baseline with Spatial attention (SA), CBAM, Non-local attention and PAM, respectively. The adoption of these attention modules can bring steady performance improvement compared to the baseline. The proposed PAM outperforms SA by +0.96% and +0.9% on mAP and Rank-1, respectively, which proves its effectiveness. By comparing several different attention modules, although Non-local attention also achieves better performance, compared to PAM, the Non-local mechanism requires more consumption. The activation map for PAM is shown in Figure 9. Compared with the activation map generated by without use PAM and by with use CBAM, our proposed pedestrian attention module learns the effective pedestrian information and suppresses irrelevant information. Compared with the general SA module, the proposed PAM has receptive fields of different dilation with shared weights, followed by the SE module, which is used to help the network adaptively learn the importance of the convolution layer with different dilation suppresses the learning of irrelevant information. As shown in Figure 9, the network can learn a more effective feature representation with the PAM.

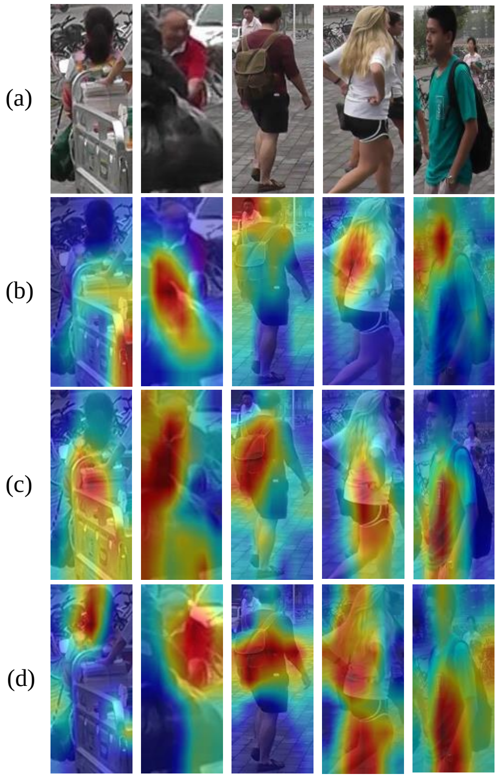


Fig. 9 Visualization of the layer 4 feature maps learned by attention module and without the attention module on PRW. From left to right, (a) Original images, (b) activation map without use PAM, (c) activation map with CBAM and (d) Our proposed PAM activation map.

Comparison of different losses. We verify the effectiveness of the proposed loss, and several losses were evaluated with the proposed method on PRW dataset. We only swap the loss functions of our method, leaving the rest part unchanged. The FC results, SphereFace, and ArcFace loss functions are shown in rows 5, 6, and 7 of Table VII. The results of the FC loss are consistently better than SphereFace loss and ArcFace loss, which shows the effectiveness of the FC loss we proposed.

V. CONCLUSION

This paper proposes a feature compactness loss (FC loss) for end-to-end person search, which learns more discriminative features of instance and improves pedestrian features' distribution without constructing a triplet. Besides, we designed a pedestrian attention module (PAM), which can effectively enhance pedestrian feature representation and suppress irrelevant information. With our proposed method, inter-identity features are more discriminative, and intra-identity features are compact. Our method achieves comparable performance on CUHK-SYSU and PRW, and extensive experiments verify the effectiveness.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3082775, IEEE Transactions on Circuits and Systems for Video Technology

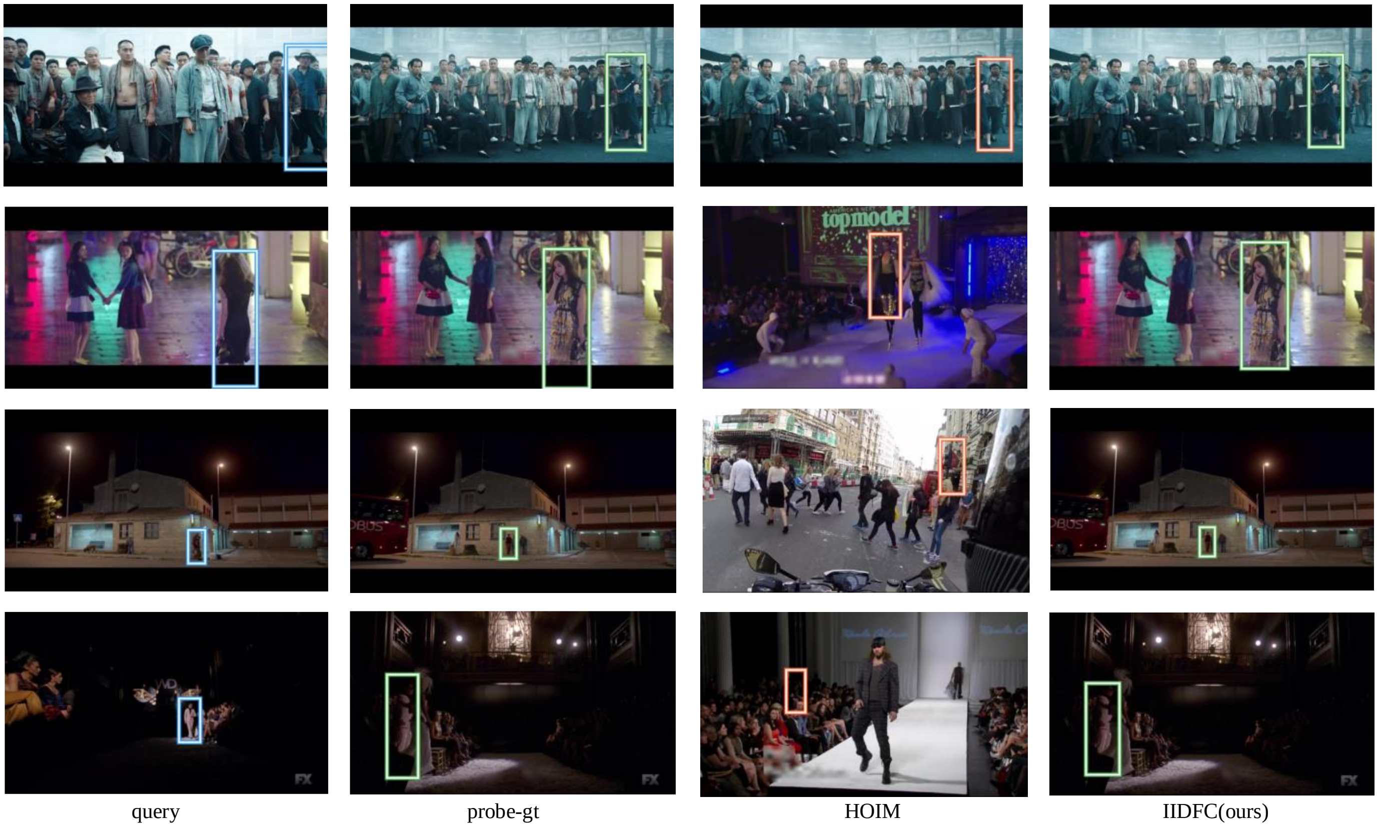


Fig. 10. Visualization of rank-1 result on CUHK-SYSU dataset. In the figure, the first column represents the image of the queried pedestrian, the second column represents the gt image of the queried pedestrian, the third column is the rank-1 search result of the baseline method HOIM, and the fourth column is the rank-1 search result of our IIDFC method.

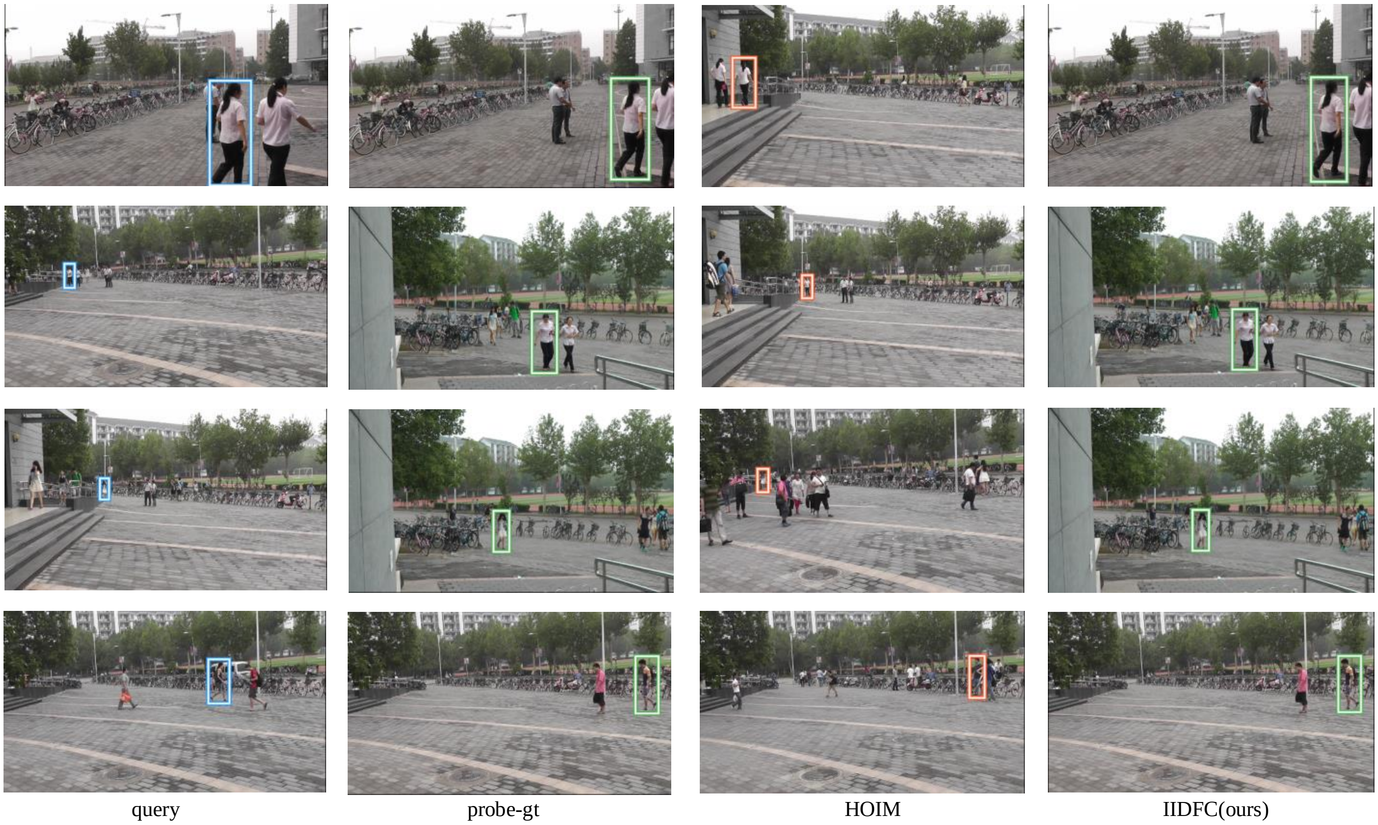


Fig. 11. Visualization of rank-1 result on PRW dataset. In the figure, the first column represents the image of the queried pedestrian, the second column represents the gt image of the queried pedestrian, the third column is the rank-1 search result of the baseline method HOIM, and the fourth column is the rank-1 search result of our IIDFC method.

REFERENCES

[1]. Xu Y, Ma B, Huang R, et al. Person search in a scene by jointly modeling people commonness and person uniqueness[C]//Proceedings of the 22nd ACM international conference on Multimedia. 2014: 937-940.

[2]. Xiao T, Li S, Wang B, et al. Joint detection and identification feature learning for person search[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 3415-3424.

[3]. Liu H, Feng J, Jie Z, et al. Neural person search machines[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 493-501.

[4]. Zheng L, Zhang H, Sun S, et al. Person re-identification in the wild[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1367-1376.

[5]. Lan X, Zhu X, Gong S. Person search by multi-scale matching[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 536-552.

[6]. Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[7]. Chang X, Huang P Y, Shen Y D, et al. RCAA: Relational context-aware agents for person search[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 84-100.

[8]. Chen D, Zhang S, Ouyang W, et al. Person search via a mask-guided two-stream cnn model[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 734-750.

[9]. Munjal B, Amin S, Tombari F, et al. Query-guided end-to-end person search[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 811-820.

[10]. Xiao J, Xie Y, Tillo T, et al. IAN: the individual aggregation network for person search[J]. Pattern Recognition, 2019, 87: 332-340..

[11]. Yan Y, Zhang Q, Ni B, et al. Learning context graph for person search[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 2158-2167.

[12]. Li J, Liang F, Li Y, et al. Fast Person Search Pipeline[C]//2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2019: 1114-1119.

[13]. Han C, Ye J, Zhong Y, et al. Re-id driven localization refinement for person search[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 9814-9823.

[14]. Dong W, Zhang Z, Song C, et al. Instance Guided Proposal Network for Person Search[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 2585-2594.

[15]. Sun Y, Zheng L, Yang Y, et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 480-496.

[16]. Zhong Y, Wang X, Zhang S. Robust Partial Matching for Person Search in the Wild[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 6827-6835.

[17]. Wang G, Yuan Y, Chen X, et al. Learning discriminative features with multiple granularities for person re-identification[C]//Proceedings of the 26th ACM international conference on Multimedia. 2018: 274-282.

[18]. Tian K, Huang H, Ye Y, et al. End-to-End Thorough Body Perception for Person Search[C]//AAAI. 2020: 12079-12086.

[19]. Chen D, Zhang S, Yang J, et al. Norm-Aware Embedding for Efficient Person Search[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 12615-12624.

[20]. Di Chen14 S Z, Ouyang W, Yang J, et al. Hierarchical Online Instance Matching for Person Search[J]. 2020.

[21]. Dai J, Zhang P, Lu H, et al. Dynamic imposter based online instance matching for person search[J]. Pattern Recognition, 2020, 100: 107120.

[22]. Dong W, Zhang Z, Song C, et al. Bi-Directional Interaction Network for Person Search[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 2839-2848.

[23]. Wang C, Ma B, Chang H, et al. TCTS: A Task-Consistent Two-Stage Framework for Person Search[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11952-11961.

[24]. Wang X, Xiao T, Jiang Y, et al. Repulsion loss: Detecting pedestrians in a crowd[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7774-7783.

[25]. Liu W, Wen Y, Yu Z, et al. Sphereface: Deep hypersphere embedding for face recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 212-220.

[26]. Yu R, Dou Z, Bai S, et al. Hard-aware point-to-set deep metric for person re-identification[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 188-204.

[27]. Bai X, Yang M, Huang T, et al. Deep-person: Learning discriminative deep features for person re-identification[J]. Pattern Recognition, 2020, 98: 107036.

[28]. Zhao, Cairong, et al. "Deep fusion feature representation learning with hard mining center-triplet loss for person re-identification." IEEE Transactions on Multimedia 22.12 (2020): 3180-3195.

[29]. Bai S, Bai X, Tian Q. Scalable person re-identification on supervised smoothed manifold[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2530-2539.

[30]. Wang Z, Ye M, Yang F, et al. Cascaded SR-GAN for Scale-Adaptive Low Resolution Person Re-identification[C]//IJCAI. 2018, 1(2): 4.

[31]. Pang Y, Xie J, Khan M H, et al. Mask-guided attention network for occluded pedestrian detection[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 4967-4975.

[32]. Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems. 2015: 91-99.

[33]. Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). IEEE, 2005, 1: 886-893.

[34]. Dollár P, Tu Z, Perona P, et al. Integral channel features[J]. 2009.

[35]. Zheng Z, Wang P, Liu W, et al. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression[C]//AAAI. 2020: 12993-13000.

[36]. Gao S, Cheng M M, Zhao K, et al. Res2net: A new multi-scale backbone architecture[J]. IEEE transactions on pattern analysis and machine intelligence, 2019.

[37]. Deng J, Guo J, Xue N, et al. Arcface: Additive angular margin loss for deep face recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 4690-4699.

[38]. Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[39]. Ren L, Lu J, Feng J, et al. Uniform and variational deep learning for RGB-D object recognition and person re-identification[J]. IEEE Transactions on Image Processing, 2019, 28(10): 4970-4983.

[40]. Rao Y, Lu J, Zhou J. Learning discriminative aggregation network for video-based face recognition and person re-identification[J]. International Journal of Computer Vision, 2019, 127(6-7): 701-718.

[41]. Chen G, Lu J, Yang M, et al. Spatial-temporal attention-aware learning for video-based person re-identification[J]. IEEE Transactions on Image Processing, 2019, 28(9): 4192-4205.

[42]. Chen W, Chen X, Zhang J, et al. Beyond triplet loss: a deep quadruplet network for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 403-412.

[43]. Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification[J]. arXiv preprint arXiv:1703.07737, 2017.

[44]. Liu H, Jie Z, Jayashree K, et al. Video-based person re-identification with accumulative motion context[J]. IEEE transactions on circuits and systems for video technology, 2017, 28(10): 2788-2802.

[45]. Zhang W, Hu S, Liu K, et al. Learning compact appearance representation for video-based person re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 29(8): 2442-2452.

[46]. McLaughlin N, del Rincon J M, Miller P. Video person re-identification for wide area tracking based on recurrent neural networks[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 29(9): 2613-2626.

[47]. Hu J, Lu J, Tan Y P. Sharable and individual multi-view metric learning[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(9): 2281-2288.

[48]. Lin C, Lu J, Wang G, et al. Graininess-aware deep feature learning for robust pedestrian detection[J]. IEEE transactions on image processing, 2020, 29: 3820-3834.

[49]. Liu W, Wen Y, Yu Z, et al. Large-margin softmax loss for convolutional neural networks[C]//ICML. 2016, 2(3): 7.

[50]. Zhao C, Lv X, Dou S, et al. Incremental Generative Occlusion Adversarial Suppression Network for Person ReID[J]. IEEE Transactions on Image Processing, 2021, 30: 4212-4224.

[51]. Liu H, Xiao Z, Fan B, et al. PrGCN: Probability prediction with graph convolutional network for person re-identification[J]. Neurocomputing, 2021, 423: 57-70.

[52]. Yao H, Xu C. Joint Person Objectness and Repulsion for Person Search[J]. IEEE Transactions on Image Processing, 2020, 30: 685-696.

[53]. Gao C, Yao R, Zhao J, et al. Structure-aware person search with self-attention and online instance aggregation matching[J]. Neurocomputing, 2019, 369: 29-38.

[54]. Zhao C, Wang X, Zuo W, et al. Similarity learning with joint transfer constraints for person re-identification[J]. Pattern Recognition, 2020, 97: 107014.

[55]. Zhao C, Chen K, Wei Z, et al. Multilevel triplet deep learning model for person re-identification[J]. Pattern Recognition Letters, 2019, 117: 161-168.

[56]. Zhao C, Chen K, Zang D, et al. Uncertainty-optimized deep learning model for small-scale person re-identification[J]. Science China Information Sciences, 2019, 62(12): 1-13.

[57]. Huang Y, Huang Y, Hu H, et al. Deeply associative two-stage representations learning based on labels interval extension loss and group loss for person re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(12): 4526-4539.

[58]. Zheng Z, Zheng L, Yang Y. Pedestrian alignment network for large-scale person re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 29(10): 3037-3045.

[59]. Ning X, Gong K, Li W, et al. Feature Refinement and Filter Network for Person Re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020.

Shaowei Hou is currently working toward a 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 search, focusing on person re-identification and person search for visual surveillance.

Cairong Zhao is a professor with the Department of Computer Science and Technology, Tongji University, Shanghai, China. He 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. His main research interests include computer vision, pattern recognition, and visual surveillance. E-mail: zhaocairong@tongji.edu.cn.

Zhicheng Chen is pursuing his master's degree at Tongji University. He received his B.E. degree in Computer Science from Shanghai University in 2019. 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.

Jun Wu (M'05–SM'14) is a professor with the Computer Science School, Fudan University, Shanghai, China. He received his BSc degree in Information Engineering and MSc degree in Communication and Electronic Systems from XIDIAN University in 1993 and 1996, respectively. He received his PhD degrees in Signal and Information Processing from the Beijing University of Posts and Telecommunications in 1999. Wu joined Tongji University as a professor in 2010. He was a Principal Scientist at Huawei and Broadcom before he joined Tongji. His research interests include Wireless Communication, Information Theory, Machine Learning, and Signal Processing.

Zhihua Wei received the B.S. and M.S. degrees from Tongji University in 2005 and 2000, respectively, and the dual Ph.D. degrees from Tongji University and Lyon2 University in 2010. She is currently a Professor with Tongji University. Her research interests include machine learning, image processing, and data mining.

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 and he serves as the Vice President of the International Rough Set Society, an Executive Manager of the Chinese Association for Artificial Intelligence, Chair of the CAAI Granular Computing Knowledge Discovery Technical Committee, a distinguished member of Chinese Computer Federation, the Vice President of the Shanghai Computer Federation, and the Vice President of the Shanghai Association for Artificial Intelligence. He serves as Associate Editor for the International Journal of Approximate Reasoning and an Editor of the Journal of Computer Research and Development (in Chinese).