IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 1, 2019, accepted September 17, 2019, date of publication September 23, 2019,
date of current version October 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2943112

Scale Voting With Pyramidal Feature Fusion Network for Person Search

ZHERAN HONG, BIN LIU, YAN LU, GUOJUN YIN, AND NENGHAI YU
School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China
Key Laboratory of Electromagnetic Space Information, Chinese Academy of Sciences, Beijing 100864, China

Corresponding author: Bin Liu (flowice@ustc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61371192, in part by the Key Laboratory Foundation of the Chinese Academy of Sciences under Grant CXJJ-17S044, and in part by the Fundamental Research Funds for the Central Universities under Grant WK2100330002.

ABSTRACT Person search aims to find the target person among a large gallery set of real scene images. Candidates should be detected and cropped before recognizing their identifications. The detection process results in a great variance of object image scales. To address this we propose Pyramidal Feature Fusion Network that integrates the top-down mid-level features to provide multi-level feature outputs. Furthermore, we propose to apply each of the mid-level features independently in ranking and we design a ranking algorithm named Scale Voting that votes among these ranking results from different feature levels to get the final ranking order. In this approach, we can make better use of the diversity and consistency information that is hidden in different mid-levels of features. The proposed algorithm achieves the state-of-the-art performance on prevalent person search datasets.

INDEX TERMS Person search, scale voting, pyramidal feature fusion network.

I. INTRODUCTION

Person search, which is firstly defined by Xu *et al.* [1], is attracting growingly more attentions for its importance in public security. The query and gallery images are complete images of real scenes such as streets corners or shopping centers rather than cropped bounding boxes of person. As shown in FIGURE 1, given a located target on the query image (yellow box), we need to not only locate but also identify the person with the same identity throughout the gallery scene images (matched object is marked with green box and other detection results are marked in red). However, in traditional person re-identification (person ReID) methods [2]–[7], which are supposed to find a certain query person in a set of gallery in the form of beforehand cropped bounding box images. From this aspect, person search makes a further attempt of applying person re-identification into more general and more realistic scenarios. In many other computer vision tasks, similarity based approaches like identifying person objects are commonly used. For example, in shot boundary detection [8], [9] (SBD), some methods evaluate the similarity distances between two selected consecutive video frames to detect the shot transitions [10]–[12].

Most recent person search algorithms [13]–[17] share the similar detection-identification streamline, as shown in FIGURE 2. The training and testing scene images are firstly fed into a detector to generate a series of detected bounding boxes that may include person targets. Then they are used to train an identification network to generate feature vectors that represent the input. The following procedure is ranking the gallery images according to their feature-wise similarity with the query, which is the same as the traditional person re-identification methods. In this manner, person search task can be divided into the detection stage and the identification stage.

Nevertheless, person search is not a simple combination of pedestrian detection and person ReID. Apart from the general challenges of person ReID, *e.g.* pose variance, illumination changes, occlusions and misalignments, person search introduces more challenges. One of the most significant challenge is that the scale of detected person images is much more variant than person images is person ReID. In real scenes, the query person and the person in the gallery may be of great scale variance. For example, a certain identity is big and clear in one image but may probably become much smaller

The associate editor coordinating the review of this manuscript and approving it for publication was Jinjia Zhou.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

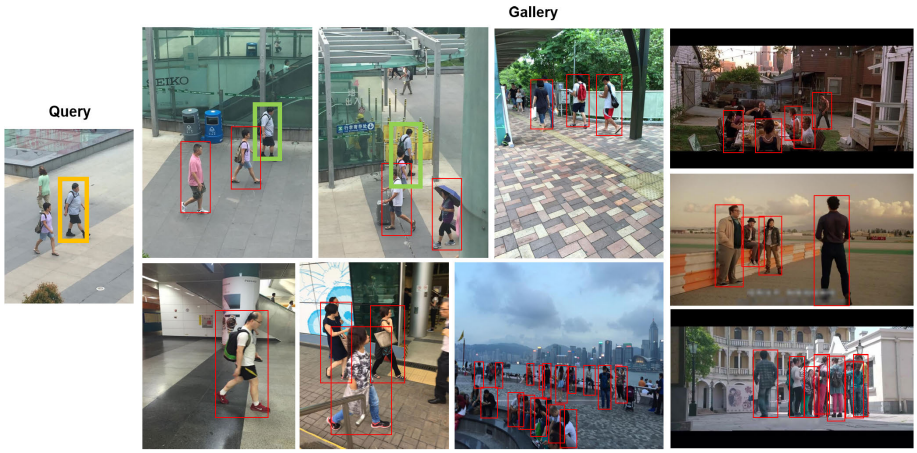


FIGURE 1. Brief overview of person search task.

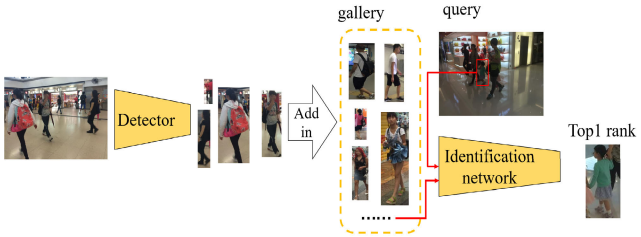


FIGURE 2. The general streamline of person search task.

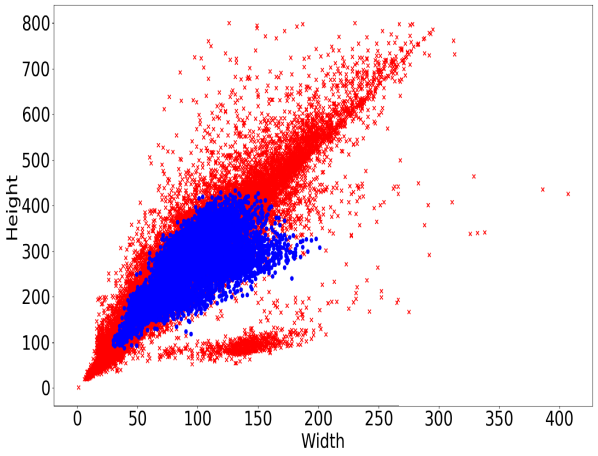


FIGURE 3. Visualization of image scale distribution of person search dataset (CUHKSYSU [16] red crosses) and person ReID dataset (CUHK03 [18] blue dots).

and more blurry in some other scenes. Xu *et al.* [15] specifically mentions the scale variance problem as the main difference of person search compared to person re-identification. In order to intuitively illustrate such difference between person ReID and person search, the image scale distribution of main stream person ReID dataset and detected person images of person search dataset are calculated and shown in FIGURE 3. The scale distribution of images in person ReID dataset CUHK03 [18] distributes closer with each other than the images from person search dataset CUHK-SYSU [16]. Since some other person ReID datasets [19], [20] releases the evenly resized images for the convenience of training and testing, there is no scale diversity contained in these datasets. In real world scenarios, the scale variation of objects is common and unpredictable. Person search task puts forward a higher requirement on the robustness of identification and detection methods. Based on the discussions above, in this paper, we develop a novel algorithm that tackles the scale variance challenge in person search.

As widely discussed in many computer vision works, shallower layer in the CNN contains richer information of small objects which tend to be merged in the high level feature maps in deeper layers. However, low-level features from shallower layers may fail to be enough representative for recognition or classification. In some classical works for semantic segmentation [21]–[23] and object detection [24], [25], they choose to apply feature fusion based structures across mid-level layers of a CNN by adding upsampled high level feature maps together with low-level feature maps. Through this top-down integration process, in low-level features, information that is ignored by high level layers can be gathered while keeping the global high-level information on the meanwhile. Similarly, we apply this multi-level feature fusion idea in the person identification structure for person search and propose **Pyramidal Feature Fusion Network (PFFN)** which generates top-down integrated feature maps to include more detailed low-level information and global high-level information.

Since the size of input images is unpredictable, it is hard and unavailable to decide which level of output feature is suitable. For blurry images with small scale, global features from the deepest layer may be enough discriminative, but for large images with rich details, features from any mid-layer may probably be of great effectiveness. Therefore, we propose to apply every mid-level features independently into the test process, and integrating the multi-level testing result via a

voting strategy. We propose **scale voting**, a ranking algorithm that provide a better ranking result among ranking orders generated from different feature levels.

However, voting method can not substantially enhance the ranking performance while given a series of littery mid-level ranking orders. Therefore, we propose to apply a series of independent triplet losses to supervise each of mid-level outputs of the network which ensures that every mid-level feature from PFFN is able to have similar recognition ability in recognizing person identities. Such training strategy is named as **multi-triplet-loss training**.

Experimental results illustrate that, all of the proposed strategies contribute to the entire system to different extents. The main contributions of this paper can de summarized in the following aspects:

1. We propose Pyramidal Feature Fusion Network (PFFN) that aims to fuse multi-level CNN feature maps in the top-down manner. And each of the mid-level feature is independently applied into the procedure of ranking the galleries.

2. Given the multi-level output features generated from the PFFN, we propose scale voting to integrate the ranking results of every mid-level features and provide a better ranking order.

3. We propose to train the PFFN with multiple triplet losses to ensure consistent representations across the mid-layers.

4. The proposed method achieves the state-of-the-art performance on prevalent person search datasets compared with other counterparts.

II. RELATED WORK

Person search methods can be separated into two main categories. First, some methods take a complete scene image as input and outputs the selected region and the corresponding similarity with the query at the same time. Xiao *et al.* [16], Chang *et al.* [26] and Liu *et al.* [27] have made attempts to realise this goal. For the dependency of detection and identification in these structures, matching problems are much more complex and difficult, which limits these methods to achieve remarkable performances. Secondly, some methods intuitively apply an object detector such as Faster RCNN [28] to locate candidate persons within all the gallery images and then feed the detection results into an identification network to find the target person. Lan *et al.* [15] proposes to use concatenated multi-level features from a deep neural network to enhance the representability of features. Chen *et al.* [13] investigates a segmentation assisted network to screen out background that included in the person image to emphasis the region of targets. The overall performance of these methods tells that the second setting of person search algorithm is straightforward and effective, so in this paper our design is based on this setting.

Ranking methods are widely explored in many computer vision tasks. Ranking methods for person re-identification are usually the strategies that further develop the information within the output features to improve the ranking results without making any changes to the feature generator itself. In Re-rank [29], a good matching pair of image is defined if they can find each other in both of their top k ranking results.

TABLE 1. Abbreviations and notations of the proposed algorithm.

network PFFN	
$C1 \sim C5$	Resnet101 backbone mid-level output layers
$P1 \sim P5$	PFFN mid-level output layers
Scale Voting Algorithm	
D	mid-level output number of PFFN
QF	query image feature set
qf_i	i-th level of query feature
GF_i	i-th level of gallery image feature sets
$gf_{i,n}$	i-th level of feature of gallery image n
GF'_i	sorted gallery feature set of level i
$l_{i,n}$	location index of gallery image n in the sorted gallery feature set GF'_i
L_n	location set of gallery image n at every feature level
idx_n	position index of gallery image n
M	number of triplets that are sampled from the dataset
m	triplet m
$f^m_{anc}, f^m_{pos}, f^m_{neg}$	feature of anchor, positive and negative image sample
$d(\cdot)$	distance metric function

In object detection or segmentation, scale variance of objects is also an important difficulty and challenge. Intuitive approach to this is to train and test the model with multi-scale input of images and concatenate the results [30]–[32]. Network architectures with top-down feature fusion mechanism [23]–[25] provide another idea of enhancing the feature-level usage on input information, which acquires outstanding performances. Training and testing the model with multiple input scales is incompatible with big training batch and is time inefficient, therefore a network with pyramidal output features might be more suitable for the person search task.

Motivated by the design of top-down feature fusion for multi-scale image feature processing in object detection and the algorithm-independent property of ranking method in person ReID, we respectively develop the PFFN network architecture to generate top-down fused multi-level features and the scale voting method that plays the role of re-rank when multiple output rankings are given. The detailed design of our algorithm is discussed in the following section.

III. ALGORITHM

Following the pipeline of state-of-the-art persons search methods, shown in FIGURE 2, the training and testing scene images are firstly fed into a detector to generate a series of detected bounding boxes that may include person targets. Then the detected target images are resized into a fixed size which is set to be 256×128. If the overlap between the detected target person and the ground truth bounding box is greater than a threshold *th*, we label this detected person object with the corresponding ground truth person ID. If the overlap of more than one detection result and a single ground truth (GT) bounding box is greater than *th*, only the highest detection result will be reserved. *th* is set to be 0.5 as in most of the published person search methods. While in the testing phase, query person and his/her location in the query scene image is given and all the detection results in the gallery scene images are included in the gallery set. A table of illustrations

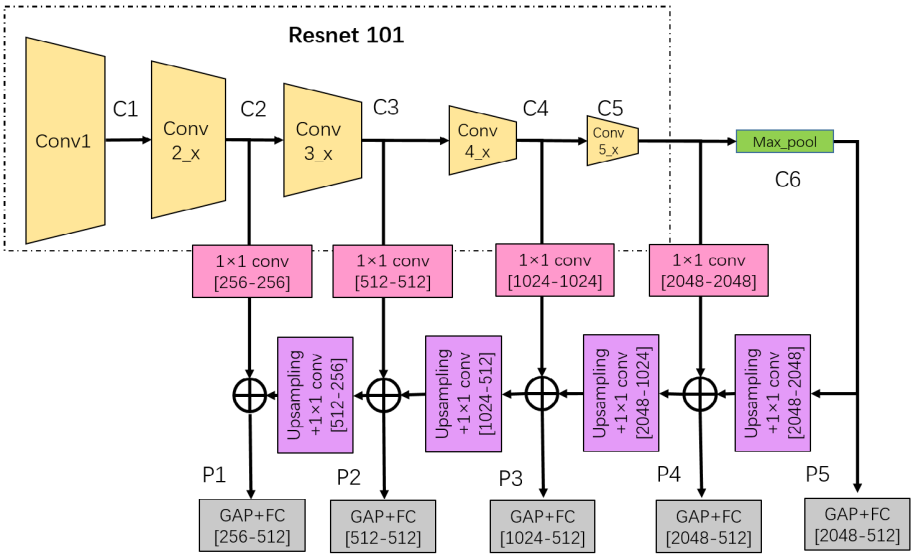


FIGURE 4. The overall architecture of Pyramidal Feature Fusion Network.

of the main abbreviations and notations in this section is given in TABLE 1.

A. PYRAMIDAL FEATURE FUSION NETWORK

The resized person images are subsequently sent into the Pyramidal Feature Fusion Network. The overall scheme of the network is shown in FIGURE 4. The mid-level features of the backbone model are fused in a series of top-down layers. Similar to the design of FPN [24], the feature maps of the deeper layers are upsampled and resized to fit the size of the shallower features maps. Different from the FPN for object detection, we concern more about the channel-wise feature representations rather than the spatial information within the feature map. We select to retain the original channel dimension of the feature maps from each level rather than regularizing them to have the same dimension of channels. Then the output feature maps with different channel dimensions are compressed into multi-level feature vectors which are further used to generate multi-level ranking results by global pooling and fully connected layers.

We use the ResNet101 [33] model as the backbone model of the proposed network. The mid-level features of the backbone are noted as C_i and $i = 1, 2, \ldots, 5$ *i.e.*, the five different scales of feature map within ResNet. For the C_1 layer is much too shallow and not representative enough, we excluded C_1 and add one more level of output C_6 by a maxpooling layer at the end of the backbone ResNet model, like [24], to generate 5 level of mid-outputs $C_2 \sim C_6$. Feature maps of such different scales are fed into an upsampling layer to make sure the feature map C_i has the same spatial size as C_{i-1}. Then the scaled feature map are shrunk or expanded in the channel dimension by a 1×1 convolutional layer to keep unanimous with C_{i-1}. Now feature maps from two adjacent levels are of the same size at every dimension. So we can combine these two feature maps with simple element-wise addition. The combined feature maps are noted by P_j and $j = 1, 2, \ldots, 5$. Through this top-down passway, more detailed information that contained in low level feature maps are fused with the global information from the topper layer. Given such multi-level feature maps, we compress them into multi-level feature vectors with coincident length L by a global average pooling layer and a fully connected layer (noted by the GAP+FC block in FIGURE 4.). Detailed information of the blocks are given in the figure, where the input and output channel dimensions of the convolutional blocks and the input and output dimensions of the FC layers are given in the following brackets, *i.e.*, $[in - out]$.

B. VARIANT SCALE VOTING WITH MULTI-LEVEL FEATURES

As we have mentioned before, for tasks with dramatic diversity of input scales, a proper single level of feature output would be difficult to choose. In the testing phase, we propose to independently apply the features from different levels to rank the gallery images then integrate the ranking results properly by a cross level multi-scale-voting strategy. Voting to keep a more generalized ranking location for each gallery image. The pipeline of this process is illustrated in FIGURE 5.

In common person search and person ReID methods, given an image of query person, all the N gallery images are sorted by their feature-level distances with the query and outputs a ranking result. In the proposed method, $D(D = 5)$ levels of feature vectors are provided by the $P1 \sim PD$ layers of the feature network, so for query image we have $QF = \{\mathbf{qf_i}|i = 1, 2, \ldots, D\}$ and for gallery images $GF_i = \{\mathbf{gf_{i,n}}|n = 1, 2, \ldots, N\}$ and $i = 1, 2, \ldots, D$, where $\mathbf{gf_{i,n}}$ is the level-i's feature vector for gallery image n. Then, we can create D levels of ranking results with the corresponding level of features

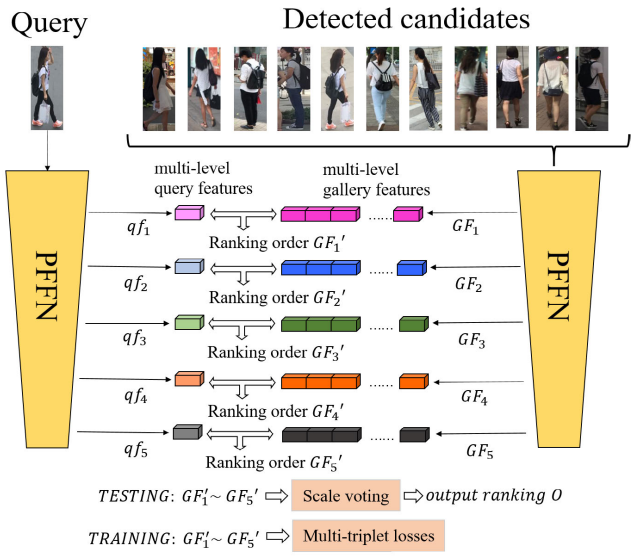


FIGURE 5. The overall pipeline of the proposed algorithms. PFFN first generates sequence of multi-level features. Then the features are subsequently applied into the multi-triplet losses for training stage and Scale voting for testing stage respectively.

of query image and gallery images. For example, we sort the N gallery images with their $\mathbf{gf_{1,n}}$ features according to their distance with $\mathbf{qf_1}$ and output the level-1 ranking order. For ranking result of level-i, the sorted images are noted as GF_i'. Then we propose variant scale voting that votes among these D ranking results: the eventual location of a certain gallery image in the final ranking order is determined when it is found in over half of the D ranking results. In this setting, gallery features of different levels vote to generate a final order, which helps to eliminate the accidental mistakes in ranking results from minor unsuitable levels while keeping the general order that exists in the majority levels. The algorithmic steps of the variant scale voting are given in Algorithm 1 in detail.

In order to better illustrate the steps of the algorithm we give a simplified example. Assume that we have 8 gallery images ($N = 8$) and 1 query image, the PFFN generate 3 level of features ($D = 3$) for each of them, as shown in TABLE 2.

Then each level of gallery features are sorted by their distance (Euclidian distance) with the query feature at the corresponding level. Then we apply variant scale voting among these three level ranking results shown in TABLE 3. For gallery image 2 ($n = 2$), colored in red, idx_2 is the location where over half of the galleries have found it. Therefore, in this example, idx_2 equals to the location where first 2 galleries have found this image in their ranking list. Thus $idx_2 = 1$ and similarly $idx_4 = 5$ (colored in blue), $idx_6 = 6$(colored in green).

Conflict occurs when more than one gallery images share the same position index idx, as shown in TABLE 4. Gallery image 2 and 5 are firstly found in the same position 2 by over half of the level of galleries, *i.e.*, $idx2 = idx5 = 2$. If such conflict happens we decide their relative order by their averaged distance with the query features across the levels. Such conflict rarely appears during our experiments.

In the proposed PFFN network, the output feature level $D = 5$ and the gallery size N differs in different datasets. With the proposed variant scale voting algorithm we achieve the goal of integrating the ranking results and keeping the predominant order within the multi-level galleries.

Algorithm 1 Variant Scale Voting With Multi-Level Features

Input: Multi-level feature vectors of a query image: $QF = \{\mathbf{qf_i} | i = 1, 2, \ldots, D\}$

Input: Ranked gallery images GF_i' and $i = 1, \ldots, D$

Output: Ranking result $O \in R^T$

```
for n = 1 ~ N (N gallery images)) do
    for i = 1 ~ D (D scales) do
        ℓ_{i,n} = index of gf_{i,n} in GF_i'
    end for
    L_n = sort[ℓ_{1,n}, ℓ_{2,n}, ..., ℓ_{D,n}]
    idx_n = the median value of L_n (Gallery image n is found by over half of the scales at idx_n)
    Labeling the gallery image n with its position index idx_n
end for
Sorting the gallery images by their idx_n and obtaining the final ranking order O
return O
```

TABLE 2. Example of query features and gallery features. Gallery size is 8 and feature level is 3.

query features	gallery features (GF)							
qf_1	$gf_{1,1}$	$gf_{1,2}$	$gf_{1,3}$	$gf_{1,4}$	$gf_{1,5}$	$gf_{1,6}$	$gf_{1,7}$	$gf_{1,8}$
qf_2	$gf_{2,1}$	$gf_{2,2}$	$gf_{2,3}$	$gf_{2,4}$	$gf_{2,5}$	$gf_{2,6}$	$gf_{2,7}$	$gf_{2,8}$
qf_3	$gf_{3,1}$	$gf_{1,2}$	$gf_{3,3}$	$gf_{3,4}$	$gf_{3,5}$	$gf_{3,6}$	$gf_{3,7}$	$gf_{3,8}$

TABLE 3. Example sorted galleries and the following scale voting operation.

query features	sorted gallery features (GF')							
qf_1	$gf_{1,2}$	$gf_{1,5}$	$gf_{1,1}$	$gf_{1,7}$	$gf_{1,4}$	$gf_{1,6}$	$gf_{1,3}$	$gf_{1,8}$
qf_2	$gf_{2,2}$	$gf_{2,1}$	$gf_{2,7}$	$gf_{2,4}$	$gf_{2,5}$	$gf_{2,6}$	$gf_{2,8}$	$gf_{2,3}$
qf_3	$gf_{3,7}$	$gf_{3,2}$	$gf_{3,4}$	$gf_{3,1}$	$gf_{3,5}$	$gf_{3,6}$	$gf_{3,3}$	$gf_{3,8}$
position	1	2	3	4	5	6	7	8

TABLE 4. Example conflict that two image have the same *idx*.

query features	sorted gallery features (GF')				
qf_1	$gf_{1,2}$	...	$gf_{1,5}$	...	...
qf_2	$gf_{2,5}$	$gf_{2,2}$	...	...	...
qf_3	...	$gf_{3,5}$	$gf_{3,2}$	...	...
position	1	2	3	4	...

C. FEATURE CONSENSUS TRAINING WITH MULTI-TRIPLET-LOSSES

Generally in person search and person ReID methods, person images are sampled to form a anchor-positive-negative (three samples) triplets. Two samples that share the same person ID are selected as the anchor and the positive sample respectively, while the negative sample is a person image with a different identity from the other two. The training batches are made up of such triplets. In [34], it is pointed out that use the hardest negative sample in the training batch as the negative sample will better benefit the model learning. Hence, we apply the triplet-loss with hard example mining [34] for our identification model training. Following is the standard triplet loss that we use:

$$L_{triplet} = \frac{1}{M}\sum_{m=1}^{M} max\{[d(f_{anc}^{m}, f_{pos}^{m}) - d(f_{anc}^{m}, f_{neg}^{m})], 0\} \quad (1)$$

where M is the triplet-level batch size and $f_{anc}^{m}, f_{pos}^{m}, f_{neg}^{m}$ respectively stand for the feature vectors of the anchor, positive sample and negative sample in triplet m. And $d(\cdot)$ means the Euclidian distance. The goal of the loss is to enlarge the distance between the samples with different identities, *i.e.*, the anchor and the negative sample, and narrow the distance between the samples of the same identity, *i.e.*, the anchor and the positive sample. When hard negative mining [34] is applied, the negative sample of a triplet is selected to be the image sample that has the greatest distance with the anchor within the triplet batch.

In common practice, the triplet loss is used to supervise a single final output of the identification network. However in the proposed method, we choose to let every mid-level feature rank independently. Only applying the supervision on a single output may result in an uncontrollable divergence of the middle features in PFFN, which will be discussed in SECTION IV. In order to address the problem and providing adaptive multi-level features for the following scale voting algorithm, every mid-level features should have a similar discriminative ability in recognizing person objects. We propose to train every mid-level output features with an independent triplet losses while training, as shown in FIGURE 5. Through this way, every middle layer learns to generate features with similar identification capacities, while laying different emphasis on the details and contexts in corresponding feature maps.

IV. EXPERIMENTS

A. EXPERIMENT SETTINGS

Experiments of the proposed method are done at CUHK-SYSU [16] dataset and PRW [35] dataset. For the detection stage we apply a Faster RCNN that is fine-tuned on the person search dataset to detect candidate person bounding boxes to build training and testing data for the following identity network.

1) CUHK-SYSU

The dataset includes 18,184 scene images with 8,432 person identities that captured from variant sources such as movies, TV shows, streetsnaps and surveillance videos. For testing set, 6,978 scene images are separated into 2,900 gallery groups at different gallery sizes.

2) PRW

This is a dataset with images captured at several fixed camera view points in campus scenes. The training set includes 5,134 scene images and testing set includes 6,112 scene images. Both sets are annotated with 34,304 bounding boxes with 932 person IDs.

The evaluation metrics of person search are top1 matching ratio (Top1) and mean average precision (mAP) which are widely used in most person search approaches [13], [15]–[17], [27]. While calculating the top1 matching ratios, best matching object should overlap the ground truth with an intersection-over-union (IoU) greater than 0.5.

The detector we used in our method is the standard Faster RCNN [28] with VGG backbone which is the same as in most recent person search methods. The Faster RCNN is only pretrained on the target person search dataset with all the labeled person bounding boxes and a ImageNet pretrained backbone. The detector is fixed during the training and testing of the identification network. We set the detection confidence threshold to be 0.7 and detect person objects which has a higher confidence than the threshold. In consideration of quantity of experimental labor and the convenience of making comparison, all of the ablation studies is done at CUHKSYSU data set with its G100 version (gallery set of 100 scene images).

B. ABLATION STUDY

In this paper, we propose a novel network architecture with top-down fusion on feature map level. Then we propose variant scale voting to integrate the results of these mid-level features. We train the entire algorithm with D independent triplet-losses to ensure the feature consistency. The ablation study is performed on the effectiveness of each of these designs.

1) ABLATION STUDY ON THE PROPOSED STAGES

We first analyse the effect of the proposed procedures in the above sections. As illustrated in TABLE 5. **m1** is the raw backbone ResNet101 model performance which is selected to be our baseline. Secondly in **m2**, we add the top-down feature integration structures in to the backbone to build the PFFN, where we train the model with single mid-level output P3 and test it on the same single output. For **m3**, we add multi-triplet loss training on all of the middle layers while test the model with single output P3. Finally, in **m4**, we train the PFFN with multi-triplet losses and test it with the proposed multi-scale voting. Note that we choose P3 layer as the single output

TABLE 5. Ablation study on the proposed stages(CUHK-SYSU G100 testing set). **m1 - m4** are the method ID.

	Resnet101	PFFN	multi-loss training	multi-scale voting	Top1(%)	mAP(%)
m1	✓				80.4	77.8
m2	✓	✓			82.9	79.1
m3	✓	✓	✓		83.4	80.2
m4	✓	✓	✓	✓	89.8	84.5

TABLE 6. Output layer testing on PFFN (CUHK-SYSU G100 testing set). Best results are marked in bold.

Output layer (TESTING)	Single output training Single output testing S-S		Multi-output training Single output testing M-S		Multi-output training Multi-scale-voting testing M-M	
	Top1(%)	mAP(%)	Top1(%)	mAP(%)	Top1(%)	mAP(%)
P1	66.6	60.4	70.1	65.7		
P2	78.3	75.3	79.0	76.7		
P3	**82.9**	79.1	**83.4**	80.2	**89.8**	**84.5**
P4	82.6	**79.3**	82.5	79.9		
P5	82.0	78.8	82.8	**80.4**		

layer for comparison since it achieves the best single-layer performances according to the following experiments (shown in TABLE 6). The top-down feature fusion structure can integrate more global and higher level features into the mid-level feature maps and mid-level features can make a complement in detailed information. This design enables the model to be more compatible with input images of variant scales. Thus the PFFN structure improves the baseline performances. Finally, the proposed scale-voting with multi-scale training strategy provides the highest improvement, which makes it reasonable to believe that ranking results from different mid-level features have the potential to generate a much better ranking result if they are properly integrated. This would be of greater help in person search task that multi-scale problem is serious.

2) COMPARISON ON SINGLE OUTPUT AND MULTI OUTPUT TRAINING AND TESTING

We further analysis the influence of output feature layer and the effectiveness of the voting strategy on multi-layer outputs. As shown in TABLE 6, we organize the training and testing under three conditions. Firstly the PFFN model is trained on a single output layer and tested on that layer (S-S). Secondly, we train all of the mid-layers of the model with multiple triplet losses while testing the output of a single layer (M-S). Lastly, the scale voting is applied instead during the testing phase after the multi-output training (M-M). The performances are given under the corresponding training conditions. Compared the results given in S-S and M-S, the multi-loss training introduces more supervision directly onto the related output layers and guides the feature fusion on each level to be more consistent. Furthermore, in the M-M column, testing results by multi-scale-voting across all of the middle layers is higher than the performance of any single layer output. This proves that there are different redundancies exist in the ranking results from different mid-level outputs. However the voting process significantly reduces such redundancy and ensures the invariant parts to remain in the final output.

TABLE 7. Example of the feature deterioration phenomenon if the model is trained without the multi-triplet-losses(CUHK-SYSU G100 testing set). The model is trained with P1 layer and tested with every single layers (P1 S-S) and scale-voting (P1 S-M).

Output layer (TESTING)	Single output training (P1) Single output testing (P1 S-S)		Single output training (P1) Multi-scale-voting testing (P1 S-M)	
	Top1(%)	mAP(%)	Top1(%)	mAP(%)
P1	66.6	60.4		
P2	5.0	1.1		
P3	0.0	0.0	0.0	0.0
P4	0.0	0.0		
P5	0.0	0.0		

Concluding the discussions in these two paragraph, the significant improvement of the proposed method is contributed by two main aspects. First is the direct supervision on every layers by multiple losses and the second is removal of the redundancies in the generated ranking orders by scale voting.

3) TRAINING WITHOUT MULTI-TRIPLET-LOSSES

As shown in TABLE 5 and TABLE 6, the improvement by multi-scale training with multiple triplet losses is not distinctive enough for the overall performance. Nonetheless, multi-triplet-losses training plays an indispensable role in preventing the multi-level features from deterioration. We give an example of the deterioration phenomenon in TABLE 7. We train the model with single output P1 and test the model with P1 and the other single output layer (P1 S-S) or test it with scale-voting (P1 S-M). An acceptable performance is achieved only when the supervised layer P1 is applied while testing (as shown in the P1 row of P1 S-S column). Testing with the other mid-layers or scale-voting among all the layers are unable to perform well under this condition. Such phenomenon is not difficult to understand since that if the model is supervised on a single output, it is hard for other parts of the network to maintain similar representative abilities. When the outputs from the mid-layers are quite different, the voting would also be more likely

TABLE 8. Ablation on PFFN (CUHK-SYSU G100 testing set) with the absence of one single middle output while multi-scale training and multi-scale voting (M-M).

Output Layers (TRAINING)					Multi-output training Multi-scale-voting testing M-M	
P1	P2	P3	P4	P5	Top1(%)	mAP(%)
✓	✓	✓	✓		85.3	82.2
✓	✓	✓		✓	85.6	82.9
✓	✓		✓	✓	84.0	81.4
✓		✓	✓	✓	86.6	84.0
	✓	✓	✓	✓	87.5	83.1
✓	✓	✓	✓	✓	89.8	84.5

to fail, that is why we get zero performance in P1 S-M. Therefore, on the purpose of retaining a consistency among all the middle outputs to be representative in identifying person objects, the multi-loss training is applied. Each of the output layer is supervised by the training objective directly so that their output rankings could be relatively more similar, which is the prerequisite for the following voting process to make improvements. In this experiment we take P1 output as example and similar results can be obtained if another mid-output is selected.

4) ABLATION ON MIDDLE LAYERS

From the above experiments, we find that P1 layer performs much poorer than the other mid-layers. It is because P1 layer is too shallow and its feature map is too large in size for being compressed into a representative feature vector by global average pooling. However, we show in TABLE 8 that each of the mid-layer (including P1) is making contribution to the final performance. First five rows are performances with one mid-level output missing and in the last row is the complete version with every mid-level applied. As indicated in the table, any single removal of the five mid-outputs would lead to a drop of the overall performance, which proves that every mid-layer output is of effectiveness.

5) EXPERIMENT ON PROPOSED PFFN AND STANDARD FPN MODEL

As we have mentioned before the top-down feature fusion structure is motivated by FPN model [24] for object detection. However we believe that person identification process in person search task depends more on the representative feature vector from the final output rather than the entire feature map. Hence we propose PFFN which generates a multi-dimensional mid-level outputs which retain the original dimension of the fused feature maps. As shown in TABLE 9, we compare the method with different base model as FPN and PFFN. The PFFN outperforms FPN in all the training and testing conditions as in the table, which proves that keeping a larger number of output channel dimensions is of significance for generating a discriminative mid-level feature vectors and proves that the PFFN model is more compatible with the proposed multi-output training and multi-scale voting.

TABLE 9. Experiment on proposed PFFN and standard FPN model. (CUHK-SYSU G100 testing set).

Method	Top1(%)	mAP(%)	Method	Top1(%)	mAP(%)
FPN(S-S)	81.3	78.7	PFFN(S-S)	82.9	79.1
FPN(M-S)	82.5	78.0	PFFN(M-S)	83.4	80.2
FPN(M-M)	86.4	82.9	PFFN(M-M)	89.8	84.5

TABLE 10. Experiment on Scale Voting with different distance metrics. (CUHK-SYSU G100 testing set).

Distance Metrics	Top1(%)	mAP(%)
Euclidian	89.8	84.5
Cosine [36], [37]	89.9	84.7
Mahalanobis(KISSME) [38]	86.6	83.1

6) EXPERIMENT ON DIFFERENT DISTANCE METRICS

In classical person ReID works, the distance metrics or metric learning methods are significant research fields. Similarly, we apply and compare some popular distance metrics in ReID to show their performances in person search task and to indicate that the proposed method is compatible with these different distance metrics. As shown in TABLE 10, Euclidian distance and Cosine distance [36], [37] achieves the similar ranking accuracy through our algorithm, while metric learning method KISSEME [38] results in a drop in the overall performance. Probably due to the scale variance challenge of person search task, a consensus covariance matrix is more difficult to learn. Metric learning methods failed to make improvements. Furthermore, to ensure fair comparisons, we apply the Euclidian distance as most published person search methods while making comparisons in subsection C.

C. COMPARISON WITH OTHER PERSON SEARCH COUNTERPARTS

We give the experimental comparisons with state-of-the-art person search counterparts and methods with detector-identifier pair that has similar design to these person search methods. Experiments are done at prevalent person search datasets CUHKSYSU and PRW. As shown in TABLE 11 and TABLE 12, methods in form $D+I+M$ represent that they are detector-identifier methods where D is the detector selected from the commonly used object detectors, I is the identifier which is selected from the classical person ReID methods, and M (if needed) is the distance metrics or metric learning approaches. The Faster RCNN [28] is written in the short form FrRCNN.

The performances of the proposed PFFN model with different training and testing conditions (S-S, M-S and M-M which are same concepts as in ablation studies) are given with the bold names in the tables. The proposed network, PFFN P3(S-S), has already achieved comparable results with many of the existing state-of-the-art methods on both datasets. In TABLE 11 we make comparisons on CUHKSYSU g100 dataset, when the proposed multi-scale

TABLE 11. Evaluation on CUHK-SYSU. Gallery size is G100. Best and second best results are marked in bold and underline respectively.

Method	Top1(%)	mAP(%)
ACF [39]+DSIFT [40]+Euclidean	25.9	21.7
ACF [39]+DSIFT [40]+KISSME [41]	38.1	32.3
ACF [39]+LOMO [6]+XQDA [6]	63.1	55.5
CCF [41]+DSIFT [40]+Euclidean	11.7	11.3
CCF [41]+DSIFT [40]+KISSME [41]	13.9	13.4
CCF [41]+LOMO [6]+XQDA [6]	46.4	41.2
CCF [41]+IDNet [16]	57.1	50.9
OIM [16]	78.7	75.5
NPSM [27]	80.1	76.3
IAN [14]	81.2	77.9
RCAA [26]	81.3	79.3
Graph [17]	<u>86.5</u>	<u>84.1</u>
FrRCNN [28]+DSIFT [40]+Euclidean	39.4	34.5
FrRCNN [28]+DSIFT [40]+KISSME [38]	53.6	47.8
FrRCNN [28]+LOMO [6]+XQDA [6]	74.1	68.9
FrRCNN [28]+IDNet [16]	74.8	68.6
FrRCNN [28]+DeepMu [42]	78.3	75.8
FrRCNN [28]+MGTS [13]	83.7	83.0
FrRCNN [28]+DPFL [43]	84.7	83.8
FrRCNN [28]+PFFN P3(S-S)	82.9	79.1
FrRCNN [28]+PFFN P3(M-S)	83.4	80.2
FrRCNN [28]+PFFN (M-M)	**89.8**	**84.5**

TABLE 12. Evaluation on PRW. Gallery size is G4000. Best and second best results are marked in bold and underline respectively.

Method	Top1(%)	mAP(%)
ACF-Alex [39]+LOMO [6]+XQDA [6]	30.6	10.3
ACF-Alex [39]+IDE$_{det}$ [35]	43.6	17.5
ACF-Alex [39]+IDE$_{det}$ [35]+CWS [35]	45.2	17.8
DPM-Alex [?]+LOMO [6]+XQDA [6]	34.1	13.0
DPM-Alex [45]+IDE$_{det}$ [35]	47.4	20.3
DPM-Alex [45]+IDE$_{det}$ [35]+CWS [35]	48.3	20.5
LDCF-Alex [46]+LOMO [6]+XQDA [6]	31.1	11.0
LDCF-Alex [46]+IDE$_{det}$ [35]	44.6	18.3
LDCF-Alex [46]+IDE$_{det}$ [35]+CWS [35]	45.5	18.3
OIM [16]	49.9	21.3
NPSM [27]	53.1	24.2
IAN [14]	61.9	23.0
FrRCNN [28]+MGTS [13]	<u>72.1</u>	<u>32.6</u>
FrRCNN [28]+PFFN P3(S-S)	57.4	21.8
FrRCNN [28]+PFFN P3(M-S)	61.2	23.6
FrRCNN [28]+PFFN (M-M)	**73.9**	**34.3**

training and testing is applied (M-M) the overall performance is distinctively improved by 6.9% in Top 1 accuracy and 5.4% in mAP. The best result exceeds the state-of-the-art by 3.3% and 0.4% in Top 1 and mAP respectively. For PRW dataset, as shown in TABLE 12, the performances of our methods remain ahead of the counterparts. The improvements of the multi-scale training and testing (M-M) is even much higher, up to 16.2% in Top1 and 11.5% in mAP, compared with the baseline (S-S). Since the number of gallery images of PRW dataset is much larger (40 times larger), the overall performances of all the methods on PRW dataset is relatively lower than that on CUHKSYSU dataset. And the greater number of gallery images, the more probable for scale variance problem to influence the ranking accuracy. That is why the proposed multi-scale approach makes more improvement on PRW dataset. The experiment results above prove that the proposed approaches have effectiveness in handling the scale variance challenge in person search. Our best result achieves 1.5% and 0.7% higher than the state-of-the-art in Top1 and mAP.

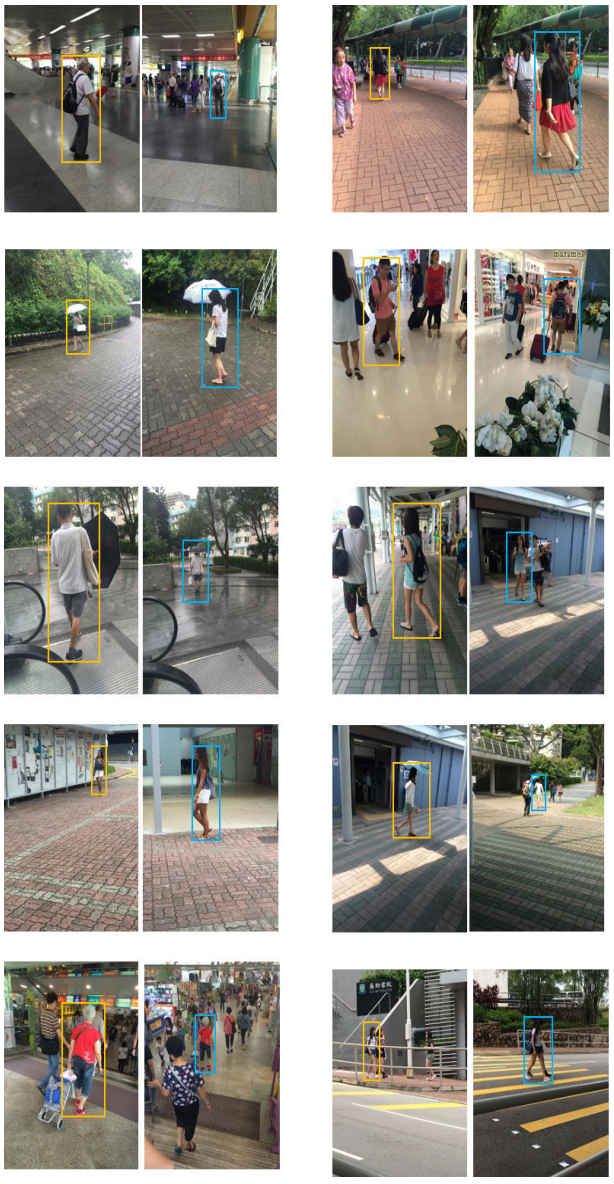


FIGURE 6. Matching pairs of the proposed method on images with significant scale variances. On the left is the query person with yellow bounding box and on the right is the top 1 matched gallery scene image and the located object in blue bounding box.

D. EXAMPLES FOR SCALE VARIANCE

In order to intuitively illustrate the performance of the proposed method on handling the scale-variance challenge in person search task, we typically select some of query and gallery pair with relative greater difference on object scales that are corrected matched by the proposed method as shown in FIGURE 6. Such examples are easy to find throughout the dataset since a great amount of the images have a

depth of scene field. In the typical cases given in the figure, the object size of the query (yellow box on the left) and the top 1 matched gallery (blue box on the right) are of great difference and our method is effective on identifying them.

V. CONCLUSION

In this paper we propose a pyramidical model PFFN with top-down feature fusion scheme, for better integrating the representative features throughout all of the middle layers of the network. Then we propose to use every mid-level output of the network structure independently in searching for the query and reorder the output ranking by a multi-scale voting strategy. In order to ensure that the output features are consensus in person representative ability, we propose to train the outputs with independent multiple triplet losses. The ablation studies that look into each steps of the proposed methods suggest that our approach is robust and effective in tackling with the multi-scale problem in person search tasks. The overall performance of the proposed method achieves a new state-of-the-art record on the prevalent challenging person search benchmarks.

REFERENCES

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

[2] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 1179–1188.

[3] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," 2018, *arXiv:1805.03344*. [Online]. Available: https://arxiv.org/abs/1805.03344

[4] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," 2018, *arXiv:1804.01438*. [Online]. Available: https://arxiv.org/abs/1804.01438

[5] R. R. Varior, M. Haloi, and G. Wang, "Gated Siamese convolutional neural network architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2016, pp. 791–808.

[6] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2197–2206.

[7] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W.-S. Zheng, "An enhanced deep feature representation for person re-identification," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–8.

[8] J. Mondal, M. K. Kundu, S. Das, and M. Chowdhury, "Video shot boundary detection using multiscale geometric analysis of nsct and least squares support vector machine," *Multimedia Tools Appl.*, vol. 77, no. 7, pp. 8139–8161, 2018.

[9] S. H. Abdulhussain, A. R. Ramli, M. I. Saripan, B. M. Mahmmod, S. A. R. Al-Haddad, and W. A. Jassim, "Methods and challenges in shot boundary detection: A review," *Entropy*, vol. 20, no. 4, p. 214, Mar. 2018.

[10] U. Gargi, R. Kasturi, and S. H. Strayer, "Performance characterization of video-shot-change detection methods," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 1, pp. 1–13, Feb. 2000.

[11] Z. Li, X. Liu, and S. Zhang, "Shot boundary detection based on multilevel difference of Colour histograms," in *Proc. 1st Int. Conf. Multimedia Image Process. (ICMIP)*, Jun. 2016, pp. 15–22.

[12] A. Adnan and M. Ali, "Shot boundary detection using sorted color histogram polynomial curve," *Life Sci. J.*, vol. 10, no. 4, pp. 1965–1972, 2013.

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

[14] J. Xiao, Y. Xie, T. Tillo, K. Huang, Y. Wei, and J. Feng, "IAN: The individual aggregation network for person search," *Pattern Recognit.*, vol. 87, pp. 332–340, Mar. 2019.

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

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

[17] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu, and X. Yang, "Learning context graph for person search," 2019, *arXiv:1904.01830*. [Online]. Available: https://arxiv.org/abs/1904.01830

[18] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 152–159.

[19] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.

[20] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2008, pp. 262–275.

[21] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Berlin, Germany: Springer, 2015, pp. 234–241.

[23] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1925–1934.

[24] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[25] K. He, G. Gkioxari, and P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

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

[27] H. Liu, J. Feng, Z. Jie, K. Jayashree, B. Zhao, M. Qi, J. Jiang, and S. Yan, "Neural person search machines," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 493–501.

[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[29] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1318–1327.

[30] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection—SNIP," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3578–3587.

[31] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2008, p. 1–8.

[32] B. Singh, M. Najibi, and L. S. Davis, "Sniper: Efficient multi-scale training," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 9310–9320.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[34] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*. [Online]. Available: https://arxiv.org/abs/1703.07737

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

[36] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2016, pp. 475–491.

[37] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, p. 0.

[38] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2288–2295.

[39] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

[40] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 3586–3593.
[41] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Convolutional channel features," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 82–90.
[42] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue, "Multi-scale deep learning architectures for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5399–5408.
[43] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2018, pp. 2590–2600.
[44] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
[45] W. Nam and P. Dollár, and J. H. Han, "Local decorrelation for improved pedestrian detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 424–432.



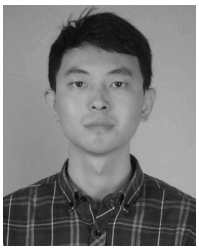
ZHERAN HONG received the B.S. degree from the Hefei University of Technology, in 2017. He is currently pursuing the M.S. degree in electrical engineering with the University of Science and Technology of China (USTC) and the Key Laboratory of Electromagnetic Space Information, Chinese Academy of Sciences (CAS). His research interests include person search and person re-identification.



BIN LIU received the B.S. and M.S. degrees from the University of Science and Technology of China (USTC), Hefei, China, in 1998 and 2001, respectively, and the Ph.D. degree from Syracuse University, Syracuse, NY, USA, in 2006, all in electrical engineering. He is currently an Associate Professor with the School of Information Science and Technology, USTC, and the Key Laboratory of Electromagnetic Space Information, Chinese Academy of Sciences (CAS). His research interests include computer vision and the Internet of Things.



YAN LU received the B.S. degree in communication engineering from Anhui Polytechnic University, in 2017. He is currently pursuing the master's degree with the University of Science and Technology of China (USTC) and the Key Laboratory of Electromagnetic Space Information, Chinese Academy of Sciences (CAS).



GUOJUN YIN received the bachelor's degree from the University of Science and Technology of China (USTC), in 2014. He is currently pursuing the Ph.D. degree with USTC and the Key Laboratory of Electromagnetic Space Information, Chinese Academy of Sciences (CAS). His research interests include computer vision, deep learning, image/video processing, natural language processing, and machine learning.



NENGHAI YU received the B.S. degree from the Nanjing University of Posts and Telecommunications, in 1987, the M.E. degree from Tsinghua University, in 1992, and the Ph.D. degree from the University of Science and Technology of China (USTC), in 2004, where he is currently a Professor. His research interests include multimedia security, multimedia information retrieval, video processing, and information hiding.