Subtask-dominated Supervised Pretraining Transfer Learning for Person Search

Chuang Liu[1]
niklaus@sjtu.edu.cn
Hua Yang[1,2]
hyang@sjtu.edu.cn
Shibao Zheng[1,2]
sbzh@sjtu.edu.cn

[1] The Institute of Image Communication and Network Engineering
Department of Electronic Engineering
Shanghai Jiao Tong University
Shanghai, China

[2] Corresponding Authors

Abstract

Since person search datasets are of limited scale due to expensive efforts to collect large-scale annotated datasets, existing one-step person methods leverage models pretrained on ImageNet to overcome this shortcoming. However, pretraining on ImageNet suffers from the large domain gap between ImageNet and target datasets. To address this issue, we propose a Subtask-dominated Supervised Pretraining (SSP) transfer learning method. The proposed SSP method takes the person re-identification (Re-ID) subtask as the dominant subtask of one-step person search and pretrains the backbone model in the Re-ID subtask with annotated data. The pretrained backbone weights can provide the one-step person search model with a better initialization to help it converge to a better solution. Specifically, the proposed SSP method surpasses the ImageNet pretraining method by 6.6% mAP and 1.9% top-1 on the PRW dataset. Besides, to reduce the impact of person detection subtask on the dominant Re-ID subtask, we further design a Multi-level RoI Fusion Pooling layer to enhance the discrimination ability of learned person features for one-step person search. Extensive experiments on the PRW and CUHK-SYSU datasets demonstrate the superiority and effectiveness of the proposed method.

1 Introduction

Person search aims to locate query persons from panoramic images. Compared to the traditional person re-identification (Re-ID) task, person search integrates person detection and person Re-ID tasks into a unified framework, making it more applicable to real-world applications. Most person search methods can be divided into two categories, the two-step methods [1, 21, 27] and the one-step methods [2, 6, 23]. Compared to the two-step framework, the one-step framework can jointly optimize the person detection and Re-ID subtasks with fewer parameters and computations. Thus, we focus on the one-step person search framework in this paper.

Model pretraining plays a crucial role in computer vision community. One typical approach is to pretrain a CNN backbone model on the large-scale ImageNet [20] dataset and transfer the pretrained backbone model to target tasks, especially when target tasks lack

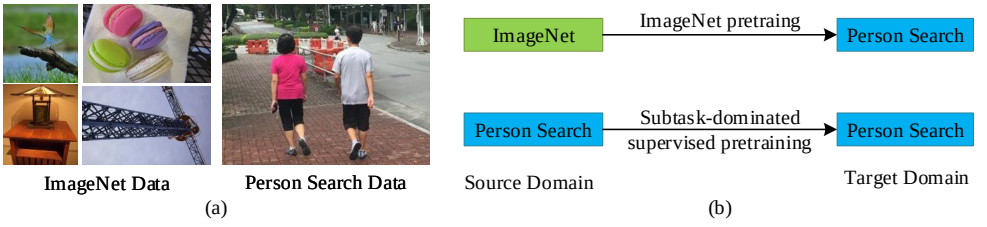

Figure 1: Comparison of ImageNet pretraining method and our proposed subtask-dominated supervised pretraining method. (a) There exists a large domain gap between ImageNet data and person search data. (b) Our proposed method can avoid the domain gap between the source domain and target domain.

enough training data. For person search, it requires great efforts to annotate a large number of person bounding boxes and corresponding person identities. As a result, existing public person search datasets (PRW [27] and CUHK-SYSU [23]) are of limited scale. To relieve the shortage of person search data, existing one-step person search methods usually use model weights pretrained on ImageNet to initialize the CNN backbone. However, it suffers from the large domain gap between ImageNet and person search data in terms of image content as shown in Fig. 1(a). In view of the recent success of self-supervised learning [4, 11], an alternative method is to conduct self-supervised pretraining for one-step person search to avoid the domain gap between the source data and the target data. Unfortunately, directly applying self-supervised pretraining to the one-step person search model does not work well due to its dependence on large-scale training data [8]. Intuitively, when the training data is of limited scale, the data annotations can effectively help to pretrain a CNN backbone capable of learning better representations, and a better pretrained backbone can also help to train a stronger one-step person search model. Therefore, a better pretraining method for person search should take advantage of the limited annotated data to pretrain the backbone.

To this end, we make the first attempt towards supervised pretraining transfer learning for one-step person search in this paper. The one-step person search has two subtasks, namely person detection subtask and Re-ID subtask. Person detection aims to distinguish the person foreground from the background by learning coarse-grained features of persons. Differently, person Re-ID aims to distinguish persons with various identities by learning fine-grained features of persons. These characteristics of the two subtasks mean it will bring little negative impact on person detection performance to give the Re-ID subtask higher priority. Besides, person detection is also not a severe bottleneck of overall person search performance [13], which means that the key to improving overall person search performance is to improve the person Re-ID subtask. Motivated by the above observations, we propose a Subtask-dominated Supervised Pretraining (SSP) transfer learning method for one-step person search.

Specifically, the proposed SSP method takes the Re-ID subtask as the dominant subtask and pretrains the backbone model in the traditional Re-ID training style directly. As shown in Figure 1(b), our SSP method pretrains the backbone directly on target person search data, and consequently avoids the domain gap. Different from the one-step person search model, the pretrained model in the traditional Re-ID training style takes as input the person Region of Interests (RoI) rather than the scene images, which makes it easier to pretrain a powerful backbone model with some effective training tricks widely-used in person Re-ID

area. Thus, the pretrained backbone model can generate discriminative person features for re-identification. Then, the weights of the pretrained backbone model are utilized to initialize the backbone of the one-step person search model to realize knowledge transfer. With a better initialization, the one-step person search model can be directly trained to learn more discriminative person features for re-identification.

Besides, to reduce the impact of the person detection subtask on the dominant person Re-ID subtask, we design a Multi-level RoI Fusion Pooling (MRFP) layer for one-step person search. Specifically, for a predicted person RoI, we propose to crop corresponding feature maps from multi-levels of the CNN parts shared by two subtasks and fuse the cropped multi-scale feature maps from different levels before feeding them to the following networks. Compared to the widely-used single-level RoI pooling operation, the proposed MRFP layer keeps more details which are helpful to distinguish different person identities for the identification networks, and consequently can help the identification networks learn more discriminative person features for re-identification.

In summary, this paper makes the following contributions to the person search community: 1) We propose the Subtask-dominated Supervised Pretraining (SSP) transfer learning method for one-step person search. The proposed SSP method not only can bridge the domain gap between the source task data and the target task data but also exploit the limited annotated data for better feature learning. 2) We design the Multi-level RoI Fusion Pooling (MRFP) layer to relieve the impact of the person detection subtask on the dominant person Re-ID subtask. The proposed MRFP layer further improves the discrimination ability of the learned person features by keeping more details about person identities for the dominant Re-ID subtask. 3) We demonstrate the superiority of our proposed method through extensive experiments on two public person search datasets.

2 Related Work

Person Search. Recently, person search has received lots of attention from computer vision community researchers. Generally, there are two mainstream person search frameworks, the two-step and the one-step.

Zheng *et al*. [27] propose a two-step framework and thoroughly evaluate combinations of different person detectors and person Re-ID models. Following the one-step framework, Chen *et al*. [1] use the Faster R-CNN [19] to detect persons and develop a two-stream CNN model to obtain representative features of persons by fusing global features and local features. Lan *et al*. [13] propose the Cross-Level Semantic Alignment to solve the multi-scale challenge by combining cross-level feature maps. Wang *et al*. [21] propose a Task-Consist Two-Stage person search framework including an identity-guided query detector to generate query-like person detections and a Detection Results Adapted Re-ID model to make the Re-ID model adapted to the detections.

Xiao *et al*. [23] propose the one-step framework based on the Faster R-CNN detection framework and design the Online Instance Matching (OIM) loss to tackle the ill-conditioned training problem. Following the one-step framework, Munjal *et al*. [17] propose a query-guided one-step person search model which can generate query-relevant proposals. Chen *et al*. [2] propose a Norm-Aware Embedding method to solve the conflict between person detection and Re-ID by disentangling person embeddings into norms and angles to conduct detection and Re-ID, respectively. Dong *et al*. [6] develop a Bi-directional Interaction Network to learn more discriminative person features by reducing the redundant information

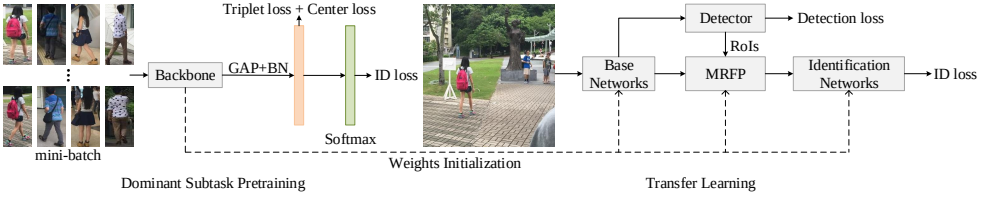

Figure 2: Pipeline of the proposed method for the one-step person search framework.

outside a bounding box.

Pretraining Transfer Learning. In person Re-ID community and even computer vision community, it is an effective and widely-used method to transfer model weights pretrained on ImageNet to target task models for better performance. However, ImageNet pretraining usually suffers from the large domain gap between ImageNet and target task datasets. Motivated by the success of self-supervised learning [3, 4, 5, 11], some person Re-ID researchers [8] propose to apply self-supervised pretraining to person Re-ID area. Despite the performance improvement, their self-supervised pretraining method [8] needs to pretrain the backbone model on the large-scale unlabeled LUPerson dataset [8] (three times larger than ImageNet-1k), which costs lots of computations and time. An alternative method is to conduct self-supervised pretraining directly on target person search data. Unfortunately, directly conducting self-supervised pretraining for person search does not work well due to its dependence on large-scale training data [8]. In this paper, we propose a subtask-dominated supervised pretraining method to avoid the above-mentioned problems and improve person search performance.

3 Method

The overall pipeline of the proposed SSP method is shown in Fig. 2. In the SSP method, the backbone is first pretrained in the dominant Re-ID subtask and then transferred to the one-step person search task. For the one-step person search model, we adopt the framework proposed by [14] with the ResNet50-IBN-a [13] as the backbone model. The backbone is divided into two parts, namely the base networks (layers from conv1 to conv3_x) and identification networks (layers from conv4_x to conv5_x). After the identification networks, a Global Average Pooling (GAP) layer followed by a Batch Normalization (BN) layer is used to obtain person features. Given an input scene image, the base networks first extract its convolutional (conv) feature maps. Then, the Faster R-CNN [19] detector predicts the possible person RoIs based on the conv feature maps from the base networks. Next, for each predicted person RoI, the proposed MRFP layer conducts multi-level multi-scale RoI feature maps cropping and fusion operations to generate the RoI feature maps for the following identification networks. Finally, the identification networks extract person features of predicted person RoIs for re-identification. In the following sections, the proposed SSP and the MRFP layer are introduced in detail.

3.1 Subtask-dominated Supervised Pretraining

Our proposed SSP method takes the Re-ID subtask as the dominant subtask of one-step person search and performs supervised pretraining for the person Re-ID subtask. Specifically, the proposed SSP method includes dominant subtask pretraining and transfer learning. Details are as follows:

Dominant Subtask Pretraining. In the one-step person search model shown in Fig. 2, networks related to the person Re-ID subtask are the base networks and identification networks. These two networks form a complete CNN backbone model. Aiming to improve the dominant Re-ID subtask, we separate the Re-ID subtask from the one-step person search task in the pretraining stage to avoid the impact of the person detection subtask. Therefore, the proposed SSP method pretrains the CNN backbone in the traditional person Re-ID training style.

Concretely, to pretrain the backbone model, a Re-ID style training dataset is first constructed by cropping all person RoIs from the original scene images in the training set. All cropped instances are resized to 256×128 ones. As depicted in Fig. 2, the pipeline proposed in [16] is employed to pretrain the backbone model. To pretrain a powerful backbone model, some useful training tricks are also adopted, including random erasing augmentation and learning rate warming up. Please kindly note that we use all the labeled person instances in the training set to conduct supervised pretraining for the CNN backbone.

Transfer Learning. To transfer the pretrained model to the one-step person search model, the weights of the pretrained Re-ID backbone are used to initialize the backbone of the one-step person search model, namely the base networks and identification networks. Although there exist scale variations between person patches and panoramic images, thanks to the local connectivity and parameter sharing of conv, conv kernels pretrained on person patches can still generate similar responses in person RoIs when applied to panoramic images to provide better initialization for the one-step person search model. Then, we retrain the one-step person search model to make it adapted to the person search task.

Following the Faster R-CNN [19], we employ the RPN training losses (L_{cls}^{rpn} and L_{reg}^{rpn}) and RoI Head training losses (L_{cls} and L_{reg}) to train the person detector. The total detection loss L_{det} is defined as follows:

$$L_{det} = L_{cls}^{rpn} + L_{reg}^{rpn} + L_{cls} + L_{reg}. \tag{1}$$

The cross-entropy loss is computed as the ID loss L_{id} to train the identification networks. Overall, the total loss to train the one-step person search model is defined as follows:

$$L = L_{id} + L_{det}. \tag{2}$$

3.2 Multi-level RoI Fusion Pooling

In the previous one-step person search model, the RoI pooling operation is conducted to crop RoI feature maps only from the output of the base networks. However, the base networks are shared by both the person detection and Re-ID subtasks. Influenced by the person detection subtask, some important details helpful to distinguish different persons are likely to be missing in the high-level outputs of the base networks, which harms the discrimination ability of person features learned by the identification networks.

Since the low-level feature maps contain more details, the proposed MRFP layer crops the RoI feature maps from multi-levels of the base networks and fuses the cropped multi-level

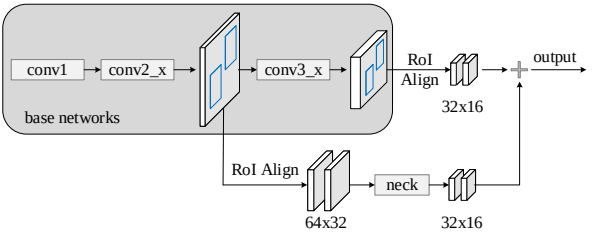

Figure 3: Illustration of the proposed MRFP layer.

multi-scale feature maps to keep more person details in the pooled feature maps. As shown in Fig. 3, based on the output feature maps of the high-level conv3_x, the RoI Align [10] pooling is conducted to pool the RoI feature maps into 32×16 ones. Besides, in the proposed MRFP layer, the RoI Align is also performed to the output feature maps of the low-level conv2_x. Since the output resolution of conv2_x is twice as large as that of conv3_x, the pooling size for conv2_x is 64×32. To perform the fusion of RoI feature maps from two levels, a neck layer is applied to transform the low-level RoI feature maps into the same size RoI feature maps as the high-level ones.

In this work, we use the conv3_x of the backbone as the neck layer to keep semantic consistency between the transformed RoI feature maps and the high-level ones. Please note that the neck layer is also initialized with the pretrained weights but does not share parameters with the conv3_x in the base networks. Then, the pixel-wise sum is utilized to fuse RoI feature maps from two levels. The fused RoI feature maps are to be fed into the identification networks to extract person features for re-identification.

4 Experiments

In this section, we run experiments on two public person search datasets, the PRW [27] and CUHK-SYSU [23], and compare the proposed method with some state-of-the-art methods. Afterwards, ablation study results are reported for each component in the proposed method.

4.1 Datasets

PRW dataset [27] is collected in Tsinghua university by six cameras. A total of 11,816 video frames containing 43,110 person bounding boxes are provided. The training set includes 5,704 frames where 15,575 person bounding boxes are labeled with 482 identities and the rest bounding boxes are unlabeled. For the testing set, 2,057 labeled person bounding boxes are marked as the query set, and 6,112 frames are taken as the gallery set. The search scope is the whole gallery set.

CUHK-SYSU dataset [23] collects video frames from the street snap and movies. A total of 18,184 frames with 96,143 person bounding boxes are provided. The training set includes 11,206 frames containing 15,080 person bounding boxes labeled with 5,532 identities and a lot of bounding boxes without identity labels. The testing set is composed of 2,900 labeled query persons and 6,978 gallery frames. Different from the PRW dataset, for each query person, the CUHK-SYSU dataset provides several gallery subsets with various gallery sizes.

4.2 Evaluation Protocols

The Cumulative Matching Characteristic (CMC top-1) and mean Average Precision (mAP) are adopted to evaluate the performance of person search. These two metrics are also the widely-used evaluation protocols in the person Re-ID area. However, different from Re-ID, the AP of each query person is scaled by its recall rate, and the mAP is calculated as the average of all APs across all query persons.

4.3 Implementation Details

For the backbone model, we pretrain it for total 120 epochs on the constructed Re-ID style training set. On the PRW dataset, we first randomly select 16 identities and choose 4 instances for each identity to obtain a balanced batch with size 64. Since the total number of identities on the CUHK-SYSU is much larger than that on the PRW dataset, we randomly select 64 identities and 4 instances for each identity to obtain a balanced batch with size 256 for better convergence during the training phase. The Adam optimizer is employed to pretrain the backbone. The initial learning rate is 3.5×10^{-4} and decayed by a factor of 10 in 40-th and 70-th epochs, respectively. The learning rate is warmed up linearly from 3.5×10^{-5} to 3.5×10^{-4} in the first 10 epochs. The weight decay factor for the Adam optimizer is 5×10^{-4}. The random horizontal flip and random erasing [16] are employed to augment the training data.

For the person search model, we train it end-to-end for total 20 epochs using the SGD optimizer with batch size 16. The learning rate is set to 0.005 initially and decayed to 0.0005 after 12 epochs. The weight decay factor for the SGD optimizer is 1×10^{-4}. Only the random horizontal flip data augmentation is used during the training of the person search model. For the RPN in the detector, the anchor sizes are set to 4, 8, 16, and 32 for each location on feature maps, and the anchor aspect ratios are set to 1, 2, and 3. The height and width of an input scene image are scaled by the same factor to make the shorter side not less than 640 pixels or the longer side not more than 960 pixels. During the reference phase, the predicted person bounding boxes with foreground scores lower than 0.5 are removed, and only bounding boxes whose Intersection over Union (IoU) with ground truth bounding boxes larger than 0.5 are regarded as true detection results.

4.4 Comparison with State-of-the-art Methods

In this section, we run experiments on the PRW and CUHK-SYSU datasets and compare the proposed person search method with some state-of-the-art methods.

Comparison on PRW. As shown in Table 1, our proposed method achieves 59.6% mAP and 89.7% top-1 accuracy, outperforming all the compared methods by large margins. In the compared one-step methods, the DKD method [26] obtains the highest performance with 50.5% mAP and 87.1 % top-1. Compared to the DKD method, our method surpasses it by 9.1% mAP and 2.6% top-1 accuracy. In the compared two-step methods, the strongest TCTS [21] method achieves 46.8% mAP and 87.5% top-1 accuracy. Compared with the TCTS method, our method obtains 12.8% mAP and 2.2% top-1 improvement. These comparison results demonstrate that our proposed method is superior to the compared state-of-the-art methods on the PRW dataset.

Comparison on CUHK-SYSU. As reported in Table 1, our proposed method achieves 93.3% mAP and 94.2% top-1 accuracy on the CUHK-SYSU dataset, surpassing most com-

	Method	PRW		CUHK-SYSU	
		mAP (%)	top-1 (%)	mAP (%)	top-1 (%)
two-step	MGTS [4]	32.6	72.1	83.0	83.7
	CLSA [13]	38.7	65.0	87.2	88.5
	RDLR [9]	42.9	70.2	93.0	94.2
	IGPN [8]	46.2	86.1	90.3	91.4
	TCTS [21]	**46.8**	**87.5**	**93.9**	**95.1**
one-step	NPSM[15]	24.2	53.1	77.9	81.2
	IAN [23]	23.0	61.9	76.3	80.1
	LCGPS [24]	33.4	73.6	84.1	86.5
	QEEPS [17]	37.1	76.7	88.9	89.1
	NAE+ [6]	44.0	81.1	92.1	92.9
	APNet [25]	41.9	81.4	88.9	89.3
	BINet [6]	45.3	81.7	90.0	90.7
	PSFL [12]	44.2	85.2	92.3	**94.7**
	DKD [26]	50.5	87.1	93.1	94.2
	AlignPS [28]	45.9	81.9	93.1	93.4
	Baseline	48.0	86.1	87.5	88.8
	Ours	**59.6**	**89.7**	**93.3**	94.2

Table 1: Comparison with the state-of-the-art methods on PRW and CUHK-SYSU.

pared methods. It is observed that TCTS [21] and PSFL [12] obtain higher performance than ours. The TCTS method is two-step, requiring two independent models (a person detector and a person Re-ID model) to tackle the person search problem. In contrast, our method is one-step and solves the person search problem in a multi-task framework with fewer computations. The PSFL method designs a Prototype Guided Attention Module for saliency feature learning in the one-step person search framework. Even if our method does not contain an attention module for saliency feature learning, it still obtains comparable top-1 performance with the PSFL method.

Moreover, experiments are also conducted to explore the influences of different gallery sizes. The gallery size ranges from 50 to 4,000. As shown in Fig. 4, the performance of all methods is degraded with the gallery size increasing, which indicates that it is still challenging to search for the target persons from the large search scope in real-world applications. When the gallery size increases, the proposed method outperforms all the compared state-of-the-art methods except a two-step method, the TCTS. Compared to the TCTS method,

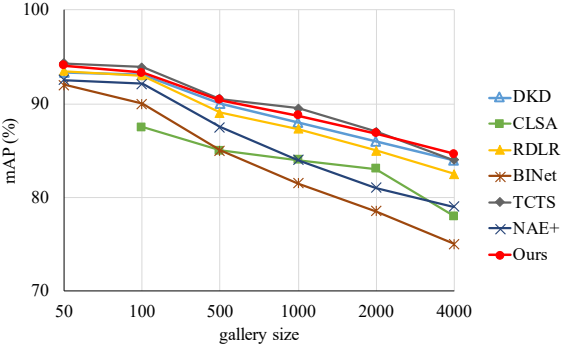

Figure 4: Comparison on the CUHK-SYSU dataset with different gallery sizes.

Method	mAP (%)	top-1 (%)
Baseline (Random init)	27.6	75.7
Baseline (ImageNet)	48.0	86.1
Baseline (Self-supervised)	46.9	85.9
Baseline (SSP)	**54.6**	**88.0**
ResNet-50 (ImageNet)	42.0	83.9
ResNet-50 (Self-supervised on LUPerson)	45.4	85.8
ResNet-50 (SSP)	**49.0**	**86.0**
OIM* (ImageNet)	42.7	84.9
OIM* (SSP)	**45.1**	**85.4**
BINet* (ImageNet)	36.7	80.1
BINet* (SSP)	**40.5**	**81.4**
Baseline (ImageNet)+MRFP	51.4	86.9
Baseline (SSP)+MRFP	**59.6**	**89.7**
ResNet-50 (ImageNet)+MRFP	45.1	86.0
ResNet-50 (Self-supervised on LUPerson)+MRFP	50.6	87.4
ResNet-50 (SSP)+MRFP	54.1	87.9

Table 2: Effectiveness of each proposed component. OIM* and BINet* represent our re-implemented models.

our method is still comparable. These experimental results demonstrate that the proposed method is more robust against gallery size variations and has advantages for large-scale person search.

4.5 Ablation Study

In this section, experiments are conducted on the PRW dataset to validate the effectiveness of each proposed component. To study the influence of each component, we replace the proposed MRFP layer in the person search framework shown in Fig. 2 with a RoI Align pooling layer to construct a baseline model. The baseline model is trained with the settings described in Section 4.3. The baseline models for ImageNet pretraining and our SSP are denoted as "Baseline (ImageNet)" and "Baseline (SSP)", respectively.

Effectiveness of SSP. In the previous person search works, the ImageNet pretraining is the most popular pretraining transfer learning method. To validate the effectiveness of the proposed SSP, we compare it with the baseline model adopting ImageNet pretraining. As shown in Table 2, the "Baseline (SSP)" model outperforms the "Baseline (ImageNet)" model by 6.6% in mAP and 1.9% in top-1. Besides, the "Baseline (SSP)+MRFP" surpasses the "Baseline (ImageNet)+MRFP" by large margins.

Moreover, we also apply the proposed SSP to some other one-step person search models to further validate its effectiveness, including the OIM model [23] and BINet [6]. As shown in Table 2, for our re-implemented OIM and BINet methods, the person search performance is further improved compared to the ImageNet pretraining method when the SSP method is applied. These experimental results further demonstrate the effectiveness of the proposed SSP method.

Effectiveness of MRFP. As shown in Table 2, person search performance is further improved when the MRFP layer is applied. Specifically, the "Baseline (SSP)+MRFP" outperforms the "Baseline (SSP)" by 5.0% in mAP and 1.7% in top-1. In addition, when the MRFP layer is applied to "Baseline (ImageNet)" and ResNet-50 based methods, person search performance is also improved, further validating the effectiveness of the MRFP layer.

Further comparison of different pretraining methods. As represented in Table 2, compared to random initialization of the CNN backbone, the ImageNet pretraining method significantly boosts person search performance thanks to the large-scale ImageNet dataset. However, there exists a large domain gap between ImageNet and target datasets, which

makes the ImageNet pretraining a suboptimal choice for the target person search task. Compared to the ImageNet pretraining, our proposed SSP method pretrains the backbone model with the annotated person search data for the dominant person Re-ID subtask, which avoids the domain gap. Consequently, our SSP method achieves higher performance. Besides the ImageNet pretraining method, we also compare the proposed SSP method with the self-supervised pretraining method. Please kindly note that the self-supervised pretraining is conducted only for the person Re-ID subtask. Table 2 shows that the performance of self-supervised pretraining is much lower than our proposed SSP. As validated in [8], the self-supervised pretraining usually requires large-scale data to learn stronger person representations. Unfortunately, the existing person search datasets are of limited scale.

Additionally, Fu *et al*. [8] propose to conduct self-supervised pretraining on the large-scale LUPerson dataset [8] for person Re-ID. We also compare our SSP with their method. Since the LUPerson dataset is three times larger than ImageNet-1k, we have no enough GPUs to pretrain a ResNet50-IBN-a backbone on the LUPerson dataset. Thus, we directly use their released ResNet-50 backbone pretrained on the LUPerson dataset. As shown in Table 2, self-supervised pretraining on the LUPerson dataset achieves better results than ImageNet pretraining. Different from ImageNet dataset, the LUPerson dataset is a person Re-ID dataset and consequently reduces the domain gap between source data and target data. Nevertheless, there is still a domain gap between the LUPerson dataset and person search datasets. Therefore, our proposed SSP obtains higher performance than self-supervised pretraining on the LUPerson. For more experimental results, please refer to the supplementary material.

5 Conclusion

In this paper, we proposed the Subtask-dominated Supervised Pretraining (SSP) transfer learning method to tackle the pretraining problem for the one-step person search. Compared to the ImageNet pretraining transfer learning method and self-supervised pretraining method, the proposed SSP method can significantly boost person search performance by reducing the domain gap between source data and target data and exploiting the annotated person search data. Besides, we further proposed the Multi-level RoI Fusion Pooling (MRFP) layer to reduce the impact of the person detection subtask on the dominant Re-ID subtask. The proposed MRFP layer can further improve the discriminative ability of the learned person features by keeping more details beneficial to the identification networks. Thorough experiments and ablation studies demonstrate the superiority and effectiveness of our proposed method.

Acknowledgment

This work was supported in part by National Natural Science Foundation of China (NSFC, Grant Nos. 62171281, 62071292), Science and Technology Commission of Shanghai Municipality (STCSM, Grant Nos. 19DZ1209303, 20DZ1200203, 18DZ2270700, 2021SHZDZX0102), and SJTU Yitu/Thinkforce Joint Laboratory for Visual Computing and Application.

References

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

[2] Di Chen, Shanshan Zhang, Jian Yang, and Bernt Schiele. Norm-aware embedding for efficient person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12615–12624, 2020.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[4] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.

[5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[6] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Bi-directional interaction network for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2839–2848, 2020.

[7] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Instance guided proposal network for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2585–2594, 2020.

[8] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. Unsupervised pre-training for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14750–14759, 2021.

[9] Chuchu Han, Jiacheng Ye, Yunshan Zhong, Xin Tan, Chi Zhang, Changxin Gao, and Nong Sang. Re-id driven localization refinement for person search. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9814–9823, 2019.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[12] Hanjae Kim, Sunghun Joung, Ig-Jae Kim, and Kwanghoon Sohn. Prototype-guided saliency feature learning for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4865–4874, 2021.

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

[14] Chuang Liu, Hua Yang, Qin Zhou, and Shibao Zheng. Making person search enjoy the merits of person re-identification. *Pattern Recognition*, 127:108654, 2022.

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

[16] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[17] Bharti Munjal, Sikandar Amin, Federico Tombari, and Fabio Galasso. Query-guided end-to-end person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 811–820, 2019.

[18] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision*, pages 464–479, 2018.

[19] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115 (3):211–252, 2015.

[21] Cheng Wang, Bingpeng Ma, Hong Chang, Shiguang Shan, and Xilin Chen. Tcts: A task-consistent two-stage framework for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11952–11961, 2020.

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

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

[24] Yichao Yan, Qiang Zhang, Bingbing Ni, Wendong Zhang, Minghao Xu, and Xiaokang Yang. Learning context graph for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2158–2167, 2019.

[25] Yichao Yan, Jinpeng Li, Jie Qin, Song Bai, Shengcai Liao, Li Liu, Fan Zhu, and Ling Shao. Anchor-free person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7690–7699, 2021.

[26] Xinyu Zhang, Xinlong Wang, Jia-Wang Bian, Chunhua Shen, and Mingyu You. Diverse knowledge distillation for end-to-end person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3412–3420, 2021.

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

[28] Yingji Zhong, Xiaoyu Wang, and Shiliang Zhang. Robust partial matching for person search in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6827–6835, 2020.