TRANSFORMER-BASED PERSON SEARCH MODEL WITH SYMMETRIC ONLINE INSTANCE MATCHING

Xuezhi Xiang[1,2], *Ning Lv*[1], *Yulong Qiao*[1,2]

[1]School of Information and Communication Engineering, Harbin Engineering University, Harbin, China
[2]Key Laboratory of Advanced Marine Communication and Information Technology,
Ministry of Industry and Information Technology, Harbin, China

ABSTRACT

Person search is a challenging retrieval problem which aims at matching pedestrians with the same identity over non-overlapping camera views. In this paper, we adopt Swin Transformer as the backbone network to extract discriminative features. We propose a symmetric online instance matching loss which transfers the symmetric idea from KL divergence to the online instance matching loss. The purpose is to strengthen the robustness of the person search model under the condition of limited training identities. We compared with the state-of-the-arts on two mainstream benchmarks: CUHK-SYSU and PRW datasets. Experimental results demonstrate the effectiveness of our method. Especially, we achieve better performance on the PRW dataset with an improvement of 6.5% and 3.5% at the mAP and top-1 accuracy, respectively.

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

1. INTRODUCTION

At present, the use of video cameras in public places is being increased in order to reduce crimes, which provide a huge amount of video surveillance resources. In security surveillance, technology is needed to identify whether pedestrians appearing in different scenes are the same person. Due to challenges such as camera resolution and shooting angle, it is difficult to retrieve pedestrians by recognizing faces. Relatively, person search technology can provide a good solution for this security task. Person search is to detect and identify pedestrians from a gallery of scene images in the case of giving images containing pedestrians as queries. A common practice is to consider person search as two subtasks: pedestrian detection and person re-identification (re-id). Compared with person re-identification, which directly uses cropped pedestrian images, person search is more close to realistic scenarios.

Relation to prior work: Along with the revival of deep learning, convolutional neural network (CNN) based methods have developed rapidly in the field of person search. According to whether features are shared by detection and re-identification, we divide existing methods into two categories: two-step methods and one-step methods.

Two-step methods [1, 2, 3, 4, 5, 6] use two separate models to handle two subtasks of person search. In these methods, typical object detectors would be directly used for pedestrian detection. More researches are being carried on revising the re-identification model and designing the training strategy to promote the connection of the two subtasks.

One-step methods [7, 8, 9, 10, 11, 12, 13] jointly optimize pedestrian detection and person re-id using a single model. The frameworks of [7, 8, 9, 10, 11, 12] are based on Faster R-CNN [14], which is an anchor-based object detector. AlignPS [13] is the first person search method based on the anchor-free detector FCOS [15]. The overall framework is jointly trained under the supervision of the detection loss and the metric learning loss such as center loss [9], Online Instance Matching (OIM) loss[7, 13], and the variants of OIM loss [10, 11, 12]. However, both OIM loss and its variants are better suited for datasets with numerous identities, ignoring the case of limited training identities.

The transformer has been widely used in the domain of natural language processing (NLP) and has achieved advanced performance in multiple tasks. Inspired by the effectiveness of transformer in NLP, researchers introduce the architecture of transformer in the field of computer vision. In [16], Swin Transformer has been proven to be superior to most CNN-based networks in processing computer recognition tasks. Hierarchical feature presentations produced by Swin Transformer are beneficial for the person search model to find pedestrians of varying scale in the scene. In this paper,

This work was supported in part by CAAI-Huawei MindSpore Open Fund, in part by the National Natural Science Foundation of China under Grant 61401113, 61871142, in part by the Natural Science Foundation of Heilongjiang Province of China under Grant LH2021F011, in part by the Fundamental Research Funds for the Central Universities of China under Grant 3072021CF0811, in part by the Key Laboratory of Advanced Marine Communication and Information Technology Open Fund under Grant AMCIT2103-03.

we utilize the tiny version of Swin Transformer as the backbone network, which is combined with multi-level feature fusion module to extract discriminative embeddings.

Inspired by the above observations, we propose an end-to-end anchor-free person search model. The contributions of this paper are summarized as follows:

(1) We utilize Swin Transformer as the backbone network. The overall network combines Swin Transformer and the CNN, which enables the model to take into account the aggregation of global information and local information.

(2) We propose the symmetric online instance matching loss for person search, which makes the model robust to the limited training identities.

(3) We conduct experiments on the CUHK-SYSU dataset and the PRW dataset. Experimental results on these two mainstream benchmarks demonstrate the effectiveness of our method compared with the state-of-the-arts.

2. METHOD

2.1. The Proposed Framework

In this work, we followed the anchor-free person search framework [13] which is based on the one-stage detector FCOS [15]. The architecture of the proposed method is depicted in Fig.1. The model contains two parts: a feature extraction module and a head module for person search.

The feature extraction module consists of a backbone network and a multi-level feature fusion network. Given input images, we adopt Swin Transformer [16] as the backbone network to extract multi-level features. The output feature maps from the stage 2, the stage 3 and the stage 4 of Swin Transformer are selected for feature fusion. We adopt the aligned feature aggregation (AFA) [13] module to process these multi-level features to obtain discriminative embeddings.

The head module for person search contains the re-id head and the detection head. In the re-id head, the fused multi-level features are directly used as re-id embeddings. We propose a symmetric online instance matching loss for optimization. For the detection head, we reserve the same network as in FCOS [15].

The architecture of backbone network is based on the tiny version of Swin Transformer (Swin-T) [16]. The multi-head self attention modules in two successive Swin Transformer blocks are with regular windowing configuration and shifted ones, respectively. In Swin-T, the window size is set to 7, and the query dimension of each head is 32. In AFA [13] module, deformable convolutional layers are used to replace conventional convolutional layers. The kernel size is set to 3. Different from the feature pyramid network, the features of different levels are integrated by concatenation, and only a single output is obtained by AFA module. The channel number of the fused feature embedding is set to 256.

2.2. Symmetric Online Instance Matching Loss

Inspired by [17], we improve the online instance matching loss (OIM) [7] according to the symmetric idea of KL-divergence. The proposed symmetric online instance matching loss (SOIM) is dedicated to strengthening the robustness of the person search model under the condition of limited training identities.

In OIM [7], the features of all labeled identities are stored in a lookup table (LUT) $V \in \mathbb{R}^{D\times L}$, and the features of unlabeled identities in the recent mini-batches are stored in the circular queue $U \in \mathbb{R}^{D\times Q}$. D is the feature dimension, L is the number of all labeled identities, and Q is the queue size. We denote the features output by the feature extraction module in a mini-batch as $x \in \mathbb{R}^D$. The probability of re-id embedding x being recognized as the identity i is calculated by softmax function as follows,

$$p_i = \frac{\exp\left(v_i^T x/\tau\right)}{\sum_{j=1}^{L}\exp\left(v_j^T x/\tau\right)+\sum_{k=1}^{Q}\exp\left(u_k^T x/\tau\right)} \quad (1)$$

where temperature parameter τ is applied to affect the final probabilities from the softmax. A low temperature (below 1) makes the model more confident. We denote the distribution of ground truth identity conditioned on feature x as q. Then, the OIM loss is defined as follows,

$$L_{OIM} = -\sum_t q_t \log p_t,\ t = 1, 2, ..., L. \quad (2)$$

In [17], the symmetric KL-divergence is described as follows,

$$SKL = KL(p||q) + KL(q||p) \quad (3)$$

Similarity, we combine the symmetric KL-divergence with the OIM loss to obtain the symmetric online instance matching loss (SOIM):

$$L_{SOIM} = L_{OIM} + L_{ROIM} = H(p, q) + H(q, p) \quad (4)$$

where $H()$ is the cross-entropy function. We define the reverse online instance matching (ROIM) loss as the reverse version of OIM.

$$L_{ROIM} = -\sum_t p_t \log(softmax(q_t)),\ t = 1, 2, ..., L. \quad (5)$$

In Eq. 5, the logarithm of the zero values could be calculated when the labeled identities are one-hot. To avoid this issue, we add the operation of softmax before the logarithm.

$$\frac{\partial L_{ROIM}}{\partial x} = -\frac{1}{\tau}\left[\left(p_t - p_t^2\right)v_t - \sum_{\substack{j=1\\ j\neq t}}^{L} p_t p_j v_j - \sum_{k=1}^{Q} p_t q_k u_k\right] \quad (6)$$

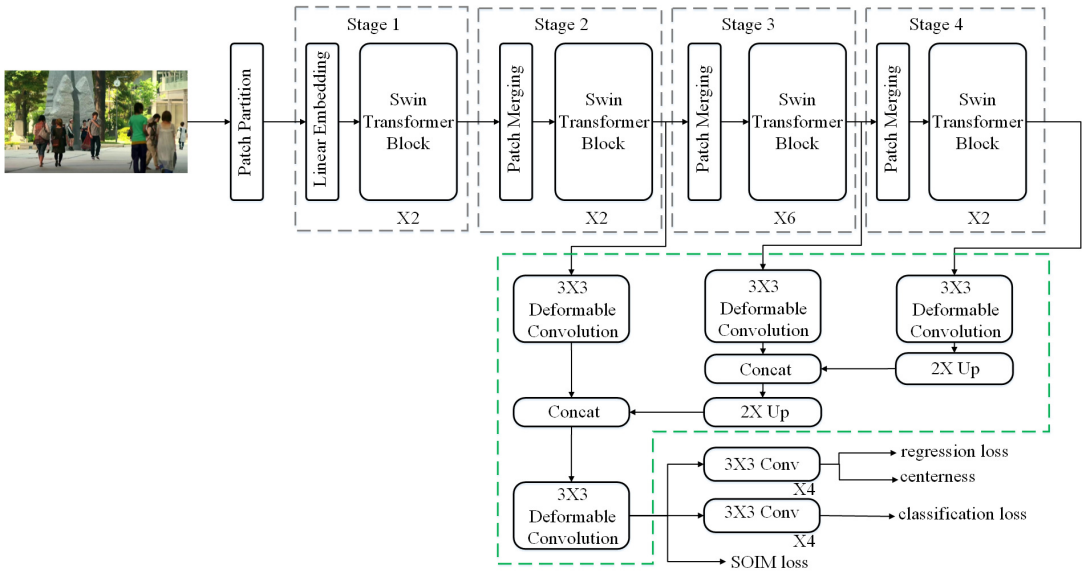

Fig. 1. The overall framework of our proposed method. The architecture of the backbone network is the tiny version of Swin Transformer (Swin-T).

Specifically, $p_t - p_t^2$ has the maximum value at $p_t = 0.5$ for $p_t \in [0, 1]$. Therefore, if the model is not confident enough about the prediction of the labeled identities, the ROIM will provide a larger acceleration to promote the sufficient learning of hard identities. As a result, the robustness of the person search model is enhanced, especially in the case of limited training identities.

Considering the difference learning paces, we treat the relative weights of the OIM loss and the ROIM loss as trainable parameters [18]. The symmetric online instance matching (SOIM) loss is defined as follows. The last two terms of SOIM are the regularizers.

$$L_{SOIM} = \frac{1}{\sigma_1^2} L_{OIM} + \frac{1}{\sigma_2^2} L_{ROIM} + \log \sigma_1 + \log \sigma_2 \quad (7)$$

In our method, the detection head and the re-id head are optimized jointly. The total loss of our method is shown as follows,

$$L_{all} = L_{SOIM} + L_{cls} + L_{reg} \quad (8)$$

L_{cls} is the focal loss [19] used to optimize the binary classification of pedestrian and background. L_{reg} is the GIoU loss [20] for bounding boxes regression which is parallel with the centerness branch.

3. EXPERIMENTS

Datasets. We conduct experiments on two person search benchmarks: CUHK-SYSU [7] and PRW [1]. The CUHK-SYSU dataset contains 18,184 images, 8,432 pedestrian identities. The training set contains 11,206 images and 5,532 identities. The test set contains 6,978 images and 2,900 pedestrian identities. The PRW dataset contains 11,816 images and 932 labeled identities totally. The training set contains 5,704 images and 482 pedestrian identities. The test set contains 6,112 images.

Evaluation protocols. We evaluate the performance of the proposed method by the standard evaluation protocols: mean average precision (mAP) and cumulative matching characteristics (CMC top-K).

Table 1. Performance comparisons with the state-of-the-arts on the CUHK-SYSU and the PRW dataset.

	Methods	Backbone	CUHK-SYSU		PRW	
			mAP(%)	top-1(%)	mAP(%)	top-1(%)
Two Step	IGPN [4]	ResNet50	90.3	91.4	47.2	87.0
	TCTS [5]	ResNet50	93.9	**95.1**	46.8	**87.5**
	OR [6]	ResNet101	93.2	93.8	52.3	71.5
One Step	OIM [7]	ResNet50	75.5	78.7	21.3	49.9
	HOIM [10]	ResNet50	89.7	90.8	39.8	80.4
	NAE+ [11]	ResNet50	92.1	92.9	44.0	81.1
	IIDFC [12]	Res2Net50	92.0	93.3	43.4	83.4
	AlignPS [13]	ResNet50	93.1	93.4	45.9	81.9
	AlinPS+ [13]	ResNet50+	**94.0**	94.5	46.1	82.1
	Ours	Swin-T	93.6	94.2	**52.4**	85.4

Implementation details. Our method is implemented on Pytorch and MindSpore platform with an NVIDIA TITAN RTX GPU. We adopt the tiny version of Swin Transformer (Swin-T) [16] pretrained on ImageNet [21] as the backbone network. During training, the multi-scale strategy is utilized where the shorter side of the image is randomly resized from 400 to 1200. The input test image are resized as 1500×900.

Table 2. Ablation study of each component

Method	Backbone	Loss	CUHK-SYSU		PRW	
			mAP(%)	top-1(%)	mAP(%)	top-1(%)
AlignPS+[13]	ResNet50+	TOIM	94.0	94.5	46.1	82.1
Ours	ResNet50+	SOIM	94.0	94.6	48.9	83.9
	Swin-T	OIM	93.6	93.9	50.4	84.3
	Swin-T	TOIM	**94.2**	**94.6**	51.5	84.1
	Swin-T	SOIM	93.6	94.2	**52.4**	**85.4**

We set the batch size to 4, the total epoch to 22 and the initial learning rate to 0.001. The learning rate is gradually warmed-up during the first 500 steps and decayed by a factor of 0.1 at epoch 16. The temperature parameter τ is the same as the setting in the code of AlignPS[1].

Comparison to the state-of-the-arts. In this subsection, we compare our proposed method with the state-of-the-arts, including two-step methods (IGPN [4], TCTS [5] and OR [6]); and one-step methods (OIM [7], HOIM [10], NAE+ [11], IIDFC [12], AlignPS and AlignPS+ [13]).

Experimental results on CUHK-SYSU dataset and PRW dataset are shown in Table 1. ResNet50+ represents the ResNet50 network with deformable convolution layers. The proposed method achieves 93.6% mAP score and 94.2% top-1 accuracy on the CUHK-SYSU dataset. On the PRW dataset, our method obtains 52.4% mAP and 85.4% top-1 accuracy, which outperforms all one-step state-of-the-arts. We adopt Swin-T, which has a complexity similar to ResNet50 [16], as the backbone to model long-range dependencies. The purpose is to explore the useful contextual information in the scene. As proved in [4, 6, 12, 13], backbone network can affect the accuracy predicted by person search model on both two datasets, which means discriminative features are important. Experimental results prove that Swin Transformer could achieve good results on person search.

In our method, we add the ROIM as an extra term to enhance the learning of the OIM on hard identities. Although the number of identities in the PRW dataset is less than that in the CUHK-SYSU dataset, the number of samples corresponding to each identity is more. The proposed SOIM loss function can effectively improve the accuracy of the model in this case. Therefore, our method achieves better results on the PRW dataset. In addition, we set the relative weights in the SOIM as trainable parameters, which is proved to be effective in the ablation experiments. Since our SOIM loss is the symmetric structure of the OIM, existing improvements on the OIM can be extended in for better performance.

Ablation study of each component. In Table 2, we consider AlignPS+ [13] as the baseline and evaluate the benefits of each proposed component on the CUHK-SYSU and the PRW datasets. On the PRW dataset, our method obtains the best performance by employing Swin-T as the backbone and SOIM as the re-id loss. Compared with ResNet50+, models with the Swin-T backbone achieve an improvement above 3.5% of mAP and 1.5% of top-1 at both the TOIM and the SOIM loss. With the same backbone, the method with SOIM can obtain higher performance on the PRW dataset. This proves that SOIM can enable more sufficient learning on hard identities of the PRW dataset. On the CUHK-SYSU dataset, we observe that the best performance is obtained by applying the Swin-T backbone and the TOIM loss. The model with SOIM does not achieve the best performance, although there is a certain improvement relative to the one with OIM. From the perspective of gradient, SOIM adds an adaptive acceleration term based on the predicted probability p_t. We analyze that the model with Swin-T is more confident about predicting the labeled identities on the CUHK-SYSU dataset and easily obtains a predicted probability $p_t > 0.5$. In this case, the acceleration term of the gradient is smaller, therefore the improvement in accuracy is not significant. Experimental results show that Swin Transformer is suitable for person search, which could extract discriminative features.

Table 3. Performance of the learnable weights in the SOIM.

Method	CUHK-SYSU		PRW	
	mAP(%)	top-1(%)	mAP(%)	top-1(%)
Ours w/o learnable weights	93.6	94.0	52.2	84.4
Ours	**93.6**	**94.2**	**52.4**	**85.4**

Performance of the learnable weights in the SOIM. In Table 3, we evaluate the performance of the learnable weights in the SOIM loss. The backbone of the model is the Swin-T. "Ours w/o learnable weights" indicates the weight of each item in the SOIM loss function is set to 1. "Ours" is the complete solution which indicates the weight of each item in the SOIM loss is learnable. Applying the learnable weights improves the performance by 0.2% on top-1 on the CUHK-SYSU dataset. On the PRW dataset, the improvement is 0.2% on mAP and 1.0% on top-1. Experimental results show that the learnable weights are beneficial to improve the performance of the model, especially on the CMC top-1 metric.

4. CONCLUSION

This paper proposes an end-to-end anchor-free person search model. The overall network combines the CNN and the transformer, which learns discriminative features for person search. In addition, we propose a symmetric online instance matching loss to strengthen the robustness of the model. The symmetric online instance matching loss is more suitable for supervising the training of the person search model with limited training identities. Experimental results on two person search benchmarks (CUHK-SYSU and PRW) show that our approach achieves performance comparable to that of the state-of-the-arts.

[1]https://github.com/daodaofr/AlignPS

5. REFERENCES

[1] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian, "Person re-identification in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3346–3355.

[2] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai, "Person search by separated modeling and a mask-guided two-stream cnn model," *IEEE Transactions on Image Processing*, vol. 29, pp. 4669–4682, 2020.

[3] C. Han, J. Ye, Y. Zhong, X. Tan, C. Zhang, C. Gao, and N. Sang, "Re-id driven localization refinement for person search," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9813–9822.

[4] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan, "Instance guided proposal network for person search," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2582–2591.

[5] C. Wang, B. Ma, H. Chang, S. Shan, and X. Chen, "Tcts: A task-consistent two-stage framework for person search," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11949–11958.

[6] Hantao Yao and Changsheng Xu, "Joint person objectness and repulsion for person search," *IEEE Transactions on Image Processing*, vol. 30, pp. 685–696, 2021.

[7] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3376–3385.

[8] B. Munjal, S. Amin, F. Tombari, and F. Galasso, "Query-guided end-to-end person search," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 811–820.

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

[10] D. Chen, S. Zhang, W. Ouyang, J. Yang, and B. Schiele, "Hierarchical online instance matching for person search," in *AAAI*, 2020.

[11] D. Chen, S. Zhang, J. Yang, and B. Schiele, "Norm-aware embedding for efficient person search," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 12612–12621.

[12] Shaowei Hou, Cairong Zhao, Zhicheng Chen, Jun Wu, Zhihua Wei, and Duoqian Miao, "Improved instance discrimination and feature compactness for end-to-end person search," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2021.

[13] Yichao Yan, Jinpeng Li, Jie Qin, Song Bai, Shengcai Liao, Li Liu, Fan Zhu, and Ling Shao, "Anchor-free person search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7690–7699.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[15] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He, "Fcos: A simple and strong anchor-free object detector," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

[16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021.

[17] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 322–330.

[18] Roberto Cipolla, Yarin Gal, and Alex Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7482–7491.

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007.

[20] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 658–666.

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.