Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.042765

ARTICLE

DTHN: Dual-Transformer Head End-to-End Person Search Network

Cheng Feng*, Dezhi Han and Chongqing Chen

School of Information Engineering, Shanghai Maritime University, Shanghai, 201306, China

*Corresponding Author: Cheng Feng. Email: 202230310105@stu.shmtu.edu.cn

Received: 11 June 2023 Accepted: 13 September 2023 Published: 31 October 2023

ABSTRACT

Person search mainly consists of two submissions, namely Person Detection and Person Re-identification (re-ID). Existing approaches are primarily based on Faster R-CNN and Convolutional Neural Network (CNN) (e.g., ResNet). While these structures may detect high-quality bounding boxes, they seem to degrade the performance of re-ID. To address this issue, this paper proposes a Dual-Transformer Head Network (DTHN) for end-to-end person search, which contains two independent Transformer heads, a box head for detecting the bounding box and extracting efficient bounding box feature, and a re-ID head for capturing high-quality re-ID features for the re-ID task. Specifically, after the image goes through the ResNet backbone network to extract features, the Region Proposal Network (RPN) proposes possible bounding boxes. The box head then extracts more efficient features within these bounding boxes for detection. Following this, the re-ID head computes the occluded attention of the features in these bounding boxes and distinguishes them from other persons or backgrounds. Extensive experiments on two widely used benchmark datasets, CUHK-SYSU and PRW, achieve state-of-the-art performance levels, 94.9 mAP and 95.3 top-1 scores on the CUHK-SYSU dataset, and 51.6 mAP and 87.6 top-1 scores on the PRW dataset, which demonstrates the advantages of this paper's approach. The efficiency comparison also shows our method is highly efficient in both time and space.

KEYWORDS

Transformer; occluded attention; end-to-end person search; person detection; person re-ID; Dual-Transformer Head

1 Introduction

Person search aims to localize a specific target person from the gallery set, which means it contains two submissions, Person Detection, and Person re-ID. Depending on these two different submissions, existing work can be divided into two-step and end-to-end methods. Two-step methods [1–6] treat them separately by conducting re-ID [7–10] on cropped person patches found by a standalone person box detector. They trade time and resource consumption for better performance, as shown in Fig. 1a.

By comparison, in a multi-task framework, end-to-end methods [11–17] effectively tackle both detection and re-ID simultaneously, as seen in Fig. 1b. These approaches commonly utilize a person detector (e.g., Faster R-CNN [18], RetinaNet [19], or FCOS [20]) for detection and then feed the feature into re-ID branches. To address the issue caused by the parallel structure of Faster R-CNN,

Li et al. [12] proposed SeqNet to perform detection and re-ID sequentially for extracting high-quality features and achieving superior re-ID performance. Yu [17] introduced COAT to solve the imbalance between detection and re-ID by learning pose/scale-invariant features in a coarse-to-fine manner and achieving improved performance. However, end-to-end methods still suffer from several challenges:

- Handing occlusions with background objects or partial appearance poses a significant challenge. The detection and correct re-ID of persons become more challenging when they are obscured by objects or positioned at the edges of the captured image. While current models may perform well in person search, they are prone to failure in complex occlusion situations.
- The significant scale of pose variations makes it complicated to re-ID. Since current models mainly utilize CNN to extract re-ID features, they tend to suffer from the scale of pose variations due to inconsistent perceptual fields, which degrades the re-ID performance.
- Efficient re-ID feature extraction remains a thorny problem. Existing methods either re-ID first or detection first, but still leave the unsolved issue of how to efficiently extract the re-ID feature for better performance.

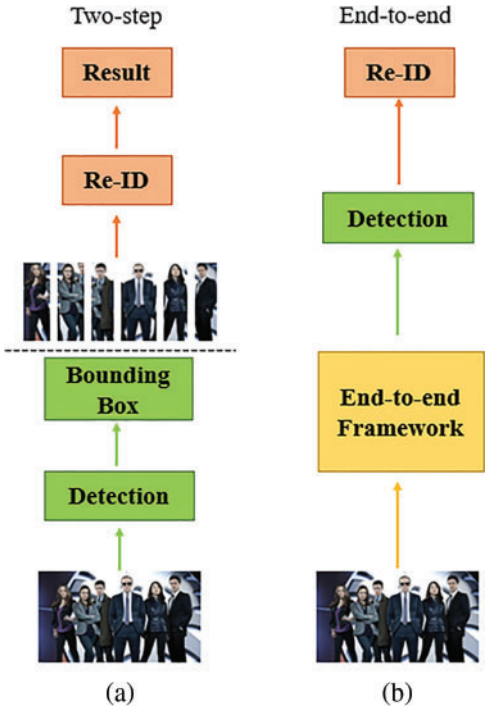

Figure 1: Classification and comparison of two person search network

For such cases, we propose a Dual-Transformer Head End-to-End Person Search Network (DTHN) method to address the above limitations. First, inspired by SeqNet, an additional Faster R-CNN head is used as an enhanced RPN to provide high-quality bounding boxes. Then a Transformer-based box head is utilized to efficiently extract box features to perform high-accuracy detection. Next, a Transformer-based re-ID head is employed to efficiently obtain the re-ID representation from the bounding boxes. Moreover, we randomly mix up partial tokens of instances in a mini-batch to learn the cross-attention. Compared to previous works that have difficulty dealing with the balance issue between detection and re-ID, DTHN can achieve high detection accuracy without degrading re-ID performance.

The main contributions of this paper are as follows:

- we propose a Dual-Transformer Head End-to-End Person Search Network, refining the box and re-ID feature extraction problem previous end-to-end frameworks were limited. The performance is improved by designing a Dual-Transformer Head structure containing two independent Transformer heads for handling high-quality bounding box feature extraction and high-quality re-ID feature extraction, respectively.
- we improve the end-to-end person search efficiency by using a Dual-Transformer Head instead of traditional CNN, reducing the number of parameters and remain a comparable accuracy. By employing the occlusion attention mechanism, the network can learn person features under occlusion, which substantially improves the performance of the re-ID in small-scale person and occlusion situations.
- we validate the effectiveness of our approach by achieving state-of-the-art performance on two widely used datasets, CUHK-SYSU and PRW. 94.9 mAP and 95.3 top-1 scores were achieved on the CUHK-SYSU dataset, and 51.6 mAP and 87.6 top-1 scores were achieved on the PRW dataset.

The remainder of this paper is organized as follows: Section 2 presents the research related to this work in recent years; Section 3 reviews the relative preparatory knowledge and presents the proposed DTHN design in detail; Section 4 presents some relevant experimental setups and verifies the effectiveness of the proposed method through experiments; Section 5 summarizes this work and provides an outlook for feature work.

2 Related Work

2.1 Person Search

Person search has received increasing attention since the release of CUHK-SYSU and PRW, two large-scale datasets. This development marked a shift in researchers' approach to person search, as they began viewing it as a holistic task instead of treating it separately. The early solutions were two-step methods, using a person detector or manually constructing the person box, then constructing a person re-ID model to search for targets in the gallery. With high performance comes high time and resource consumption, two-step methods tend to consume more computational resources and time to perform at the same level as end-to-end methods. End-to-end person search has attracted extensive interest due to the integrity of solving two submissions together. Li et al. [12] shared the stem representations of person detection and re-ID, solving two submissions sequentially. Yan [14] proposed the first anchor-free person search method to address the misalignment problem at different levels. Furthermore, Yu [17] presented a three-cascade framework for progressively balancing person detection and re-ID.

2.2 Vision Transformer

Transformer [21] was initially designed to solve problems in natural language processing. Since the release of Vision Transformer (ViT) [22], it has become popular in computer vision (CV) [23–26]. This pure Transformer backbone achieves state-of-the-art performance on many CV problems and has been shown to extract multi-scale features that traditional CNNs struggle with. The re-ID process heavily relies on fine-grained features, making it a promising technology in this field. Several efforts have been made to explore the application of ViT in person re-ID. Li et al. [27] proposed the part-aware Transformer to perform occluded person re-ID through diverse part discovery. Yu [17] performed the person search with multi-scale convolutional Transformers, learning discriminative re-ID features and distinguishing people from the background in a cascade pipeline. Our paper proposes

a Dual-Transformer Head for the end-to-end person search network to efficiently extract high-quality bounding boxes feature and re-ID feature.

2.3 Attention Mechanism

The attention mechanism plays a crucial role in the operation and function of the whole Transformer. After the proposal of ViT, numerous variants of ViT have tried to bring different features to the Transformer by changing the attention mechanism. Among them, in the target detection task, using a combination of artificial token transformations has become a mainstream approach to solve the detection of occluded targets. Based on this, Yu [17] proposed an occlusion attention module in which both positive and negative samples in the same mini-batch are randomly partially swapped to simulate the encountered background occlusion of a person, achieving good performance. This is also mainly the attention mechanism used in this paper.

To give the reader further insight into the work in this paper, Table 1 provides a summary of the related work and the work in this paper.

Table 1: A summary of related person search works and our work

Work	Summary
Person search	Person search can be seen as a combination of pedestrian detection and person re-ID. It has broad application prospects in video surveillance, finding lost children, self-service supermarkets, etc.
Two-step methods	Two-step methods conduct person re-ID on cropped person patches found by a separate object detector.
End-to-end methods	End-to-end methods jointly solve the detection and re-ID sub-problems in a more efficient, multi-task learning framework.
Vision Transformer	Based on the original Transformer model for natural language processing, Vision Transformer (ViT) is the first pure Transformer network to extract features for image recognition.
Attention mechanism	The attention mechanism plays a crucial role in Transformers. The proposed occluded attention module considers token cross-attention between either positive or negative instances from the mini-batch.
Our proposed	We propose a Dual-Transformer Head for the end-to-end person search network, refining the box and re-ID feature extraction problem previous end-to-end frameworks were limited.

3 Methods

As previously mentioned, existing end-to-end person search works still struggle with the conflict of person detection and person re-ID. Prior studies have indicated that, despite a potential decrease in detection precision, the precision of re-ID can be maintained or even improved through serialization. However, achieving a high-level detection precision results in accurate bounding box features, which are beneficial for re-ID. Thus, we propose the Dual-Transformer Head Person Search Network (DTHN) manage to get both high-quality detection and refined re-ID accuracy.

3.1 End-to-End Person Search Network

As shown in Fig. 2, our network is based on the Faster R-CNN object detector backbone with Region Proposal Network. We start by pre-processing the image to be searched for which will be converted to a size of $800 * 1500$ as a standard input. We then use the ResNet-50 [28] backbone to extract the 1024-dim backbone feature in a size of $1024 * 58 * 76$, then fed it into the RPN to obtain the region proposals. During training, RoI-Align is performed using the proposals generated by RPN to obtain the features of the region of interest for bounding box search, but RoI-Align is performed using a Ground-truth bounding box during the re-ID phase. Note that instead of using ResNet-50 stage 5 (res5) as our box head, we utilize a Transformer to extract high-quality box features and get high detection accuracy, and use the predictor head of Faster R-CNN to obtain high-confidence detection boxes. The RoI-Align operation is applied to pool a $h * w$ region as our region of interest, we use it as the stem feature $F \in R^{h*w*c}$. Note that F has the height of h and the width of w, and c denotes the number of channels. We set the intersection-over-union (IoU) thresholds at 0.5 in the training phase to distinguish positive and negative samples, and 0.8 IoU in the testing phase to get high-confidence bounding boxes. Then a Transformer re-ID head is utilized to extract distinguish features from the F. In each Transformer head, we learn the feature supervised by two losses L_{reg1} and L_{reg2}. Where N_p denotes the number of positive samples, r_i denotes the calculated regression of i-th positive samples, Δ_i denotes the corresponding ground truth regression, and L_{loc} denotes the Smooth-L_1-Loss. The expressions for L_{reg1} and L_{reg2} are identical, as shown in the equation for L_{reg} below.

$$L_{reg} = \frac{1}{N_p} \sum_{i=1}^{N_p} L_{loc}(r_i, \Delta_i) \tag{1}$$

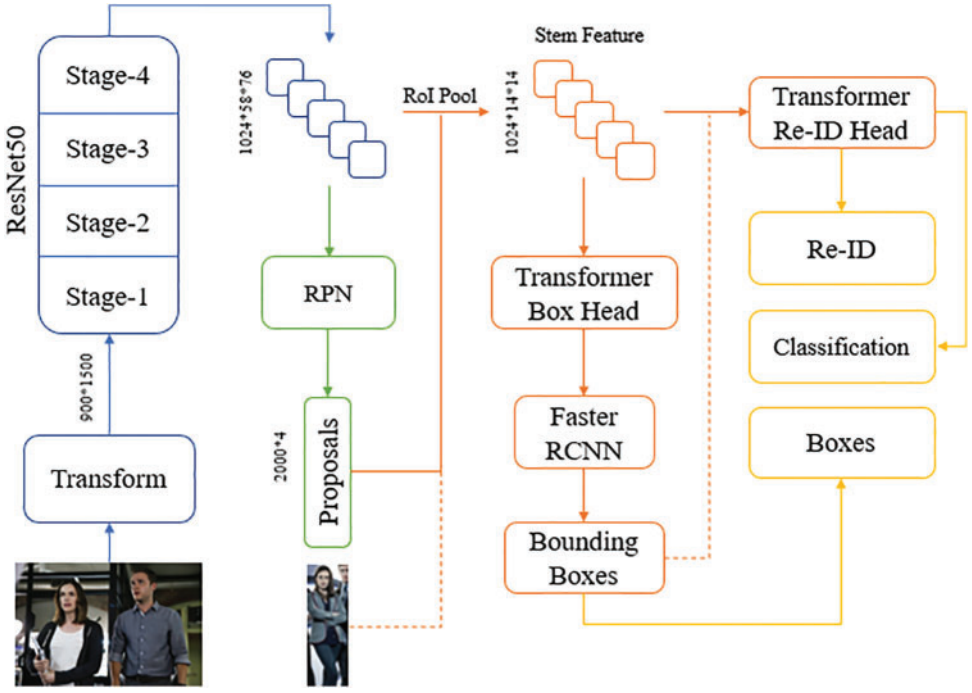

Figure 2: Structural framework of the DTHN, the dotted line means only happens in the testing phase

In addition, we also calculate the classification loss L_{cls1}, and L_{cls2} after two transformer heads. Where N denotes the number of samples, p_i denotes the predicted classification probability of i-th sample, and c_i denotes the ground truth label.

$$L_{cls1} = -\frac{1}{N}\sum_{i=1}^{N} c_i log\,(p_i) \tag{2}$$

Note that L_{cls2} and the re-ID loss L_{reid} are two different losses calculated by the Norm-Aware Embedding (NAE) L_{nae} (.), where f denotes the extracted 256-dim features.

$$L_{cls2}, L_{reid} = L_{nae}\,(f) \tag{3}$$

Then we define the overall loss function, where λ_i denotes the weight of each loss.

$$\mathrm{L} = \lambda_1 L_{reg1} + \lambda_2 L_{cls1} + \lambda_3 L_{reg2} + \lambda_4 L_{cls2} + \lambda_5 L_{reid} \tag{4}$$

3.2 Occluded Attention

The attention mechanism plays a crucial role in the Transformer. In our application, where we aim to extract high-quality bounding boxes and re-ID features, we must address the issue of occlusion. To this end, we use occluded attention in the DTH to prompt the model to learn the occlusion feature and address it in real applications, as shown in Fig. 3. Equations should be flushed to the left of the column.

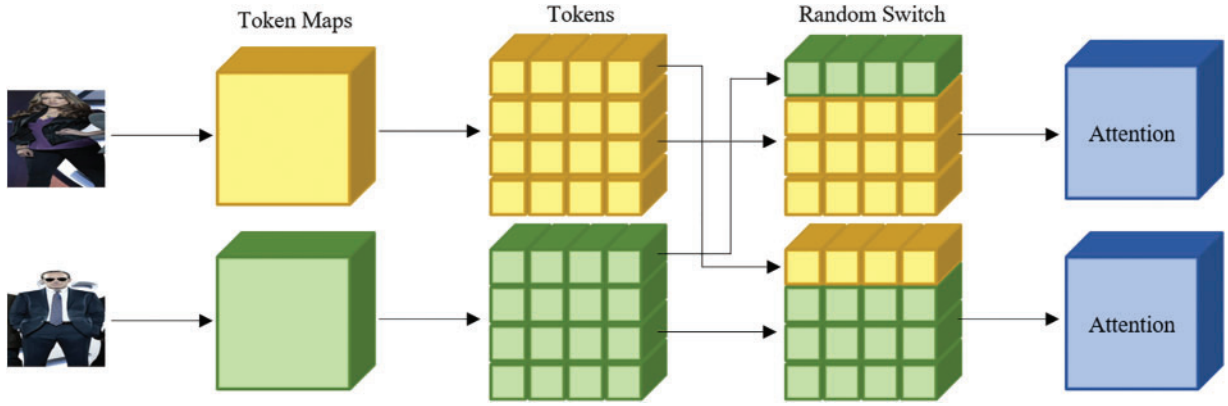

Figure 3: The occluded attention mechanism in DTHN

First, we build the token bank $X = \{x_1, x_2, \ldots, x_p\}$, where p denotes the number of box proposals, and x_i denotes the token in one mini-batch. We exchange part of the tokens with another token from the token bank according to the index, using Token-Mix-Up (TMU) function, where x_i and x_j denote the token to be handled, R denotes the random value generated by the system, T denotes the exchange threshold.

$$TMU\,(x_i, x_j) \tag{5}$$

$$when\ R > T \tag{6}$$

After random swapping, we transform the tokenized features into three matrices through three fully connected (FC) layers: query matrix Q, key matrix K and value matrix V, and then we compute the multi-head self-attention (MSA) as follows, where $\hat{c}$ denotes the channel scale of the token, it equals $\frac{c}{n}$, n is the number of slices during tokenization, m denotes the number of heads MSA has:

$$\text{MSA}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{\hat{c}/m}}\right)V \tag{7}$$

After MSA, we perform Feed Forward Network (FFN) to output features for feature regression, classification, and re-ID.

3.3 Dual-Transformer Head

The Dual-Transformer Head (DTH) consists of two individual Transformer heads designed for detection and re-ID. Although working in different parts of the network, the detection and re-ID heads share the same mechanism. The Transformer box head takes box proposals as input and generates processed features as output. In contrast, the Transformer re-ID head takes ground truth as input during the training phase but proposals during the testing phase. Therefore, we hypothesize that the quality of detection can positively impact the re-ID performance. To provide a visual representation, the structure of the DTH is visualized in Fig. 4.

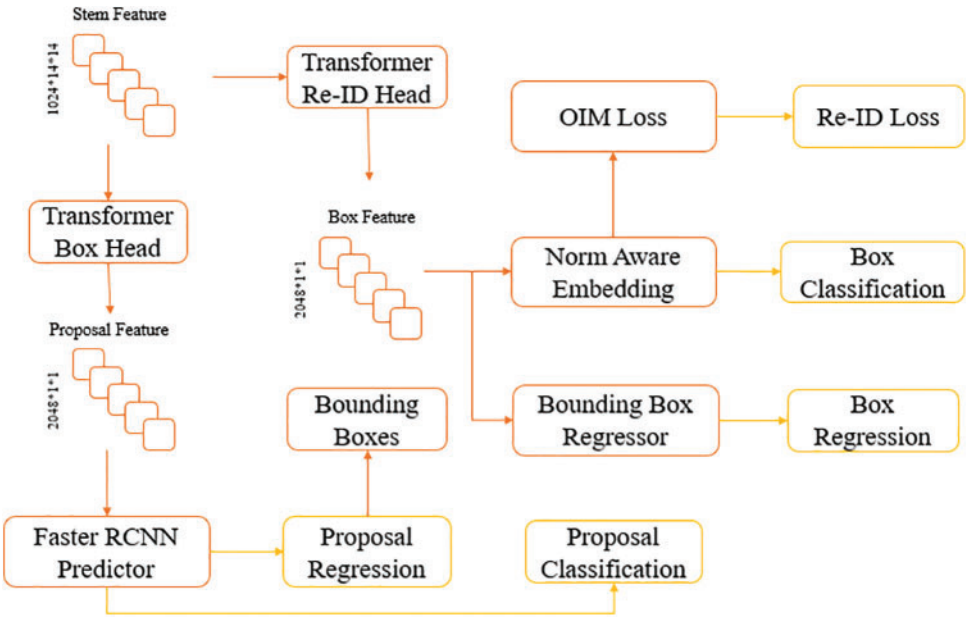

Figure 4: The structure of DTH and how it works

First, the pooled stem feature $F \in R^{h*w*c}$ is fed into the Transformer box head and obtains the proposal feature, which is fed into Faster R-CNN to calculate the proposal regression and proposal classification. After that, F is re-fed into the Transformer re-ID head and obtains box feature, which is fed into the bounding box regressor and Norm-Aware Embedding to calculate the box regression

and box classification. The loss function of NAE to calculate the box classification L_{cls2} is shown in equation below:

$$L_{cls2} = -ylog\left(\widetilde{r}\right) - (1-y)\,log\left(1-\widetilde{r}\right) \tag{8}$$

$$where\widetilde{r} = \sigma\left(\frac{r - E\left[r\right]}{\sqrt{Var\left[r\right] + \epsilon}} \cdot \gamma + \beta\right) \tag{9}$$

where $y \in \{0, 1\}$ denotes that the box is a person or background. *norm* $r \in [0, \infty)$. σ denotes the sigmoid activation function, within which is a batch normalization layer. The OIM loss is calculated using the features processed by NAE. OIM only consider the labeled and unlabeled identities, while leave the other proposals untouched. OIM has two auxiliary structures, Look-Up Table (LUT) to store all feature vectors with tagged identities and Circular Queue (CQ) to store untagged identities detected in the recent mini-batch. Based on these two structures, the probability of x being recognized as the identity with class-id i and the i-th unlabeled identity by two Softmax function. OIM loss is calculated as equation below as our re-ID loss.

$$\mathrm{p_i} = \frac{\exp\left(\mathrm{v_i^T x}/\tau\right)}{\sum_{\mathrm{j=1}}^{\mathrm{L}} \exp\left(\mathrm{v_j^T x}/\tau\right) + \sum_{\mathrm{k=1}}^{\mathrm{Q}} \exp\left(\mathrm{u_k^T x}/\tau\right)} \tag{10}$$

$$q_i = \frac{exp\left(u_i^T x/\tau\right)}{\sum_{j=1}^{L} exp\left(v_j^T x/\tau\right) + \sum_{k=1}^{Q} exp\left(u_k^T x/\tau\right)} \tag{11}$$

$$L_{reid} = E_x\left[logp_t\right] \tag{12}$$

where v_i^T denotes the i-th column of the LUT, u_i^T denotes the i-th column of the CQ, τ denotes softer probability distribution, E_x denotes the expectation, p_t denotes the probability of being judged as t.

We take the Transformer re-ID head as an example to demonstrate the process. After the feature has been pooled into $F \in R^{h*w*c}$, F will go through the tokenization. We split F to n slices channel-wise getting $\overline{F} \in R^{h*w*\hat{c}}$. We utilize series convolutional layers to generate tokens based on $\overline{F}$ getting $\hat{F} \in R^{\hat{h}*\hat{w}*\hat{c}}$. By flattening $\hat{F}$ into token $x \in R^{\hat{h}\hat{w}*\hat{c}}$. After finishing TMU, go through the MSA and FFN mentioned above transforming each token to enhance its representation ability. The enhanced feature will be projected into the same size it gets in, $\hat{h} * \hat{w} * \hat{c}$. Then we concatenate the features of the n scales of transformers to the original size $h * w * c$. There is a residual connection outside each transformer. After the global average pooling (GAP) layer, the feature Transformer outputs will be pooled and delivered to different loss functions according to the type of Transformer head. The internal structure of the Transformer head is shown in Fig. 5.

4 Experiment

All training processes are conducted in PyTorch with one NVIDIA A40 GPU, while testing processes are conducted with one NVIDIA 3070Ti GPU. The origin image will go through the ResNet-50 stage 4 and be resized to $900 * 1500$ as the input. The source code and implementation details can be found in https://github.com/FitzCoulson/DTHN/tree/master.

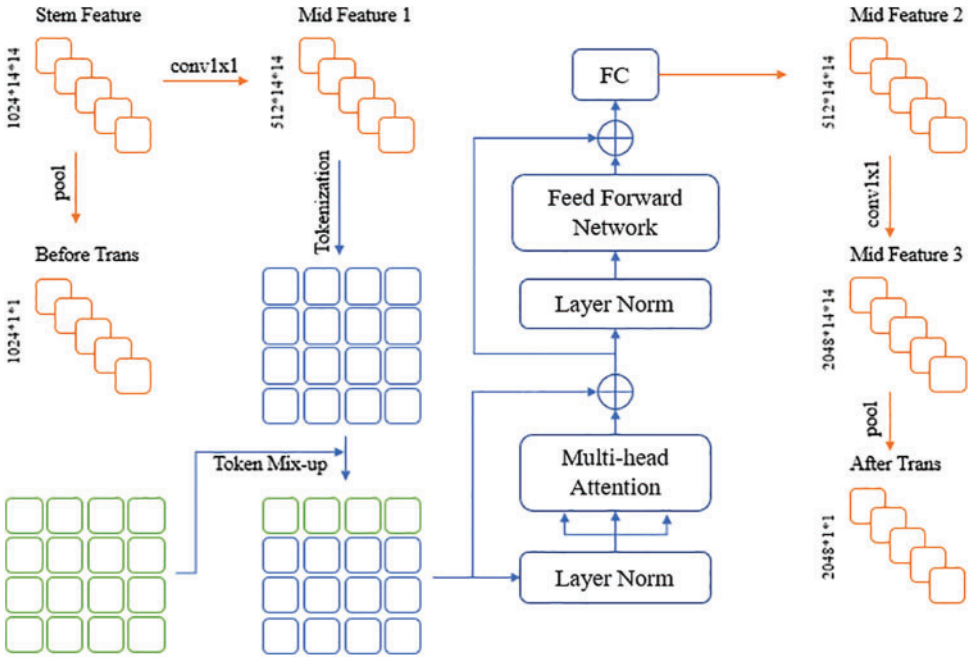

Figure 5: The internal structure of Transformer head

4.1 Datasets and Metrics

We conduct our experiments on two wildly used datasets. The CUHK-SYSU dataset [13] contains images from 18184 scenes with 8432 identities and 96143 bounding boxes. The default gallery contains 2900 testing identities in 6978 images with a default size of 100. While the PRW dataset [6] collects 11816 video frames from 6 cameras with 5704 frames and 482 identities, dividing into a training set with 5705 frames and 482 identities and a testing set with 2057 query persons in 6112 frames.

We evaluate our model following the standard evaluation metrics. According to the Cumulative Matching Characteristic (CMC), the detection box will only be considered correct when the IoU is more than 0.5. So, we use Recall and Average Precision (AP) as the performance metric for person detection. While the person re-ID uses the mean Average Precision (mAP) and top-1 scores. All the metrics the higher the better.

$$\text{mAP} = \frac{\sum_n (R_n - R_{n-1})\, P_n}{C} \tag{13}$$

where R_n and P_n separately denote the recall and precision of the n-th confidence threshold, C denotes the number of all classifications. The top-1 score denotes the result with the highest accuracy under the classification.

4.2 Implementation Detail

We take ResNet-50 pre-trained on the ImageNet as the backbone. The batch size is set to 5 during training and 1 during testing. The size of the F will be set to $14 * 14 * 1024$. The number of heads m in MSA is set to 8. The loss weight λ_1 is set to 10, and others are set to 1. We use the SGD optimizer with a momentum of 0.9 to train 20 epochs. The initial learning rate will warm up to 0.003 during the first

epoch and decrease by 10 after the 16th epoch. The CQ size of OIM is set to 5000 for CUHK-SYSU and 500 for PRW. The IoU threshold is set to 0.4 in the testing phase.

4.3 Ablation Study

We conducted several experiments on the PRW dataset to analyze our proposed method. As shown in Table 2, we test several combinations of different box heads and re-ID heads and evaluate their performance on the PRW dataset.

Table 2: Comparison with different heads

Box head	Re-ID head	Recall	AP	mAP	top-1
ResNet-50 (stage 5)	ResNet-50 (stage 5)	96.28	93.91	46.68	83.42
Transformer	ResNet-50 (stage 5)	95.71	93.32	46.26	82.74
ResNet-50 (stage 5)	Transformer	96.13	93.69	50.67	84.39
Transformer	Transformer	95.75	93.39	50.72	85.08

We set the default box head and re-ID head as ResNet-50 (stage 5) and conduct one experiment, follow by two experiments by setting the box head or the re-ID head to the corresponding Transformer head, respectively, and finally set both the box head and the re-ID head to the Transformer head for one experiment. As we can see from Table 2, when using ResNet-50 (stage 5) as the box head and the re-ID head, both detection and re-ID are at a moderate level. However, when we change the box head to Transformer, the detection accuracy does not improve, while the re-ID accuracy is also slightly reduced, so Transformer cannot play a good effect only for the box head. When we maintain the box head as ResNet-50 (stage 5), and replace the re-ID head with Transformer, the re-ID accuracy increases significantly, which shows that Transformer can maximize information extracted from the feature for re-ID. Finally, we replace both the box head and re-ID head with Transformer, while the detection accuracy is slightly reduced, the re-ID accuracy is significantly improved with the support of the DTH. As can be seen, although the Transformer box head reduces the detection accuracy, it efficiently extracts the valid information and improves the overall re-ID performance with the Transformer re-ID head. The Transformer re-ID head undoubtedly enhances the re-ID performance in various occlusion scenarios, and significantly increases the overall re-ID performance.

Therefore, we believe that our design of the DTHN can fully extract both the box features and the unique features of the person for efficient re-ID.

4.4 Comparison with State-of-the-Art Models

We compare our DTHN with state-of-the-art methods on CUHK-SYSU and PRW, including two-step and end-to-end methods. The results are shown in Table 3.

Context Bipartite Graph Matching (CBGM) is a algorithm used in test phase to integrate context information into the matching process. It compares the two most similar targets and use K-M algorithm to the optimal matching with largest weight.

Table 3: Comparison with SOTA models

	Methods	CUHK-SYSU		PRW	
		mAP	top-1	mAP	top-1
Two-step	MGTS [1] (2018)	83.0	83.7	32.6	72.1
	CLSA [4] (2018)	87.2	88.5	38.7	65.0
	RDLR [3] (2020)	93.0	94.2	42.9	70.2
	IGPN [2] (2020)	90.3	91.4	**47.2**	87.0
	TCTS [5] (2020)	**93.9**	**95.1**	46.8	**87.5**
End-to-end	IAN [29] (2019)	76.3	80.1	23.0	61.9
	NPSM [30] (2017)	77.9	81.2	24.2	53.1
	RCAA [11] (2018)	79.3	81.3	-	-
	CTXG [15] (2019)	84.1	86.5	33.4	73.6
	QEEPS [31] (2019)	88.9	89.1	37.1	76.7
	HOIM [32] (2020)	89.7	90.8	39.8	80.4
	APNet [16] (2020)	88.9	89.3	41.9	81.4
	BINet [33] (2020)	90.0	90.7	45.3	81.7
	NAE [34] (2020)	91.5	92.4	43.3	80.9
	NAE+ [34] (2020)	92.1	92.9	44.0	81.1
	DMRNet [35] (2021)	93.2	94.2	46.9	83.3
	PGS [36] (2021)	92.3	**94.7**	44.2	85.2
	AlignPS [14] (2021)	93.1	93.4	45.9	81.9
	AlignPS+ [14] (2021)	94.0	94.5	46.1	82.1
	SeqNet [12] (2021)	93.8	94.6	46.7	83.4
	AGWF [37] (2021)	93.3	94.2	53.3	**87.7**
	COAT [17] (2022)	94.2	94.7	53.3	87.4
	ROI-AlignPS [38] (2023)	**94.5**	**95.2**	50.7	84.0
	SAT [39] (2023)	94.4	94.8	**54.5**	87.5
	DTHN (ours)	93.9	94.3	50.7	85.1

Note: The bold font indicates the currently best results among the two methods.

The results of using CBGM are shown in Table 4.

Table 4: Comparison with SOTA models using CBGM

	Methods	CUHK-SYSU		PRW	
		mAP	top-1	mAP	top-1
End-to-end	AlignPS+ [14] + CBGM [12]	94.2	94.3	46.9	85.7
	SeqNet + CBGM [12]	94.8	**95.7**	47.6	87.6
	COAT [17] + CBGM [12]	94.8	95.2	**54.0**	**89.1**
	DTHN + CBGM (ours)	**94.9**	95.3	51.6	87.6

The graphical representations of each dataset's results are shown in Figs. 6 and 7. The horizontal axis is mAP and the vertical axis is top-1.

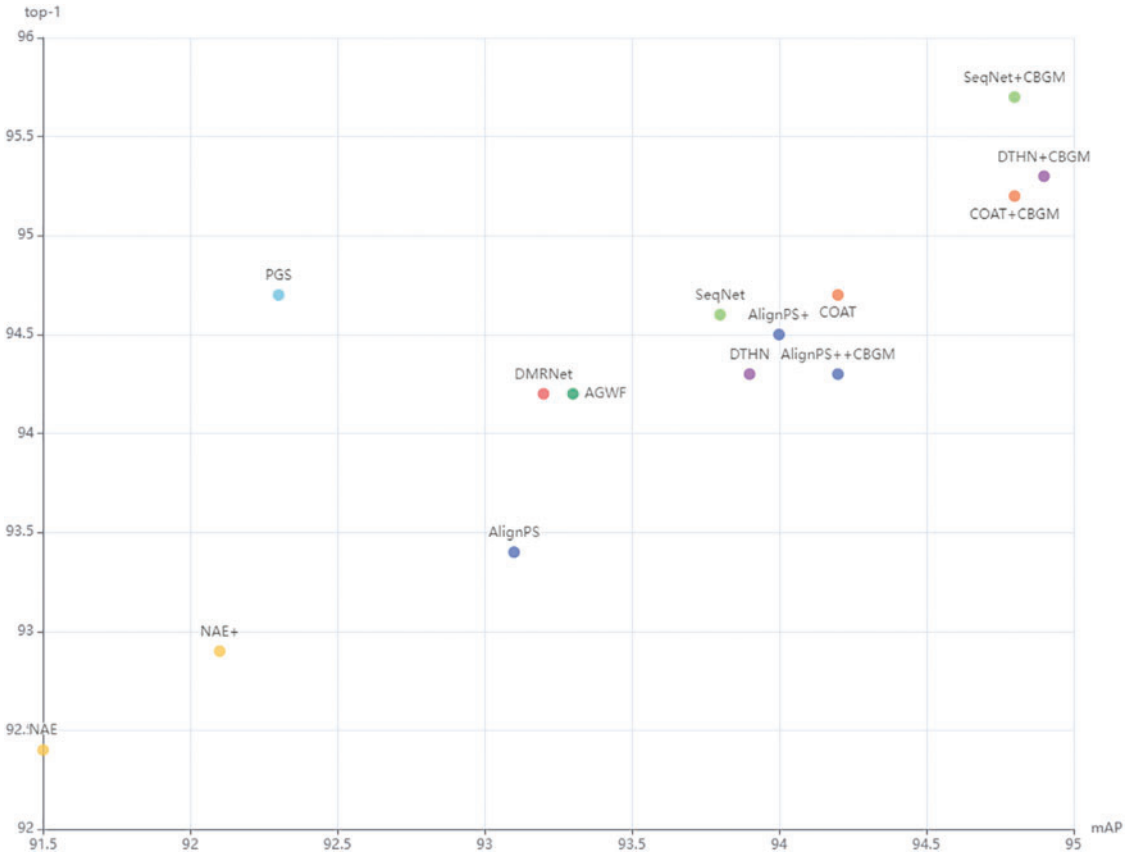

Figure 6: Comparison with SOTA end-to-end models in CUHK-SYSU

4.4.1 Result on CUHK-SYSU

As shown in the table, we achieved the same 93.9 mAP and a comparable 94.3 top-1 scores compared to the state-of-the-art two-step method TCTS. Compared with the recent end-to-end works, our mAP outperforms the AlignPS, SeqNet, and AGWF, and our top-1 score outperforms the AlignPS and AGWF. Additionally, by using the post-processing operation CBGM, both mAP and top-1 scores of our method improved to 94.9 and 95.3, achieving the best mAP in all methods with a highly competitive top-1 scores.

4.4.2 Result on PRW

PRW dataset is well known as more challenging. We achieved 50.7 mAP and 85.1 top-1 scores. Our mAP outperforms all the two-step methods. Among the end-to-end methods, our mAP and top-1 score outperform AlignPS and SeqNet, while remaining a 2.5 gap with AGWT and COAT. Due to the structural advantage of COAT, it remains state-of-the-art status on the PRW dataset, but the DTHN proposed in this paper still achieves respectable results with a smaller number of parameters and computational effort. However, by applying CBGM as a post-processing operation, we obtain a slight gain of 0.9 mAP and a significant gain of 2.5 for the top-1 score, further improving the performance

of our method and reducing the gap with COAT. This means that our proposed DTHN is effective in handling the challenging PRW dataset.

Figure 7: Comparison with SOTA end-to-end models in PRW

4.4.3 Efficiency Comparison

We compare our efficiency with two end-to-end networks SeqNet and COAT. All experiments are conducted on the RTX 3070Ti GPU on the PRW dataset. As shown in Table 5, we include the number of parameters, the multiply-accumulate operations (MACs), and the running speed in frames per second (FPS) in the comparison.

Table 5: Efficiency comparison

Methods	Param (M)	MACs (G)	FPS	mAP	top-1
SeqNet	23.90	77.23	9.43	46.7	83.4
COAT	17.48	78.33	6.79	53.3	87.4
DTHN (ours)	13.00	77.37	6.85	50.7	85.1

Compared with SeqNet and COAT, we significantly reduce the number of parameters and remain the equivalent MACs, achieving a comparable accuracy. In terms of FPS, SeqNet has the highest

9.43 because it does not need to compute attention, and we have a slight advantage in running speed compared to COAT with also computes attention. In summary, our model can run efficiently while having a good performance.

4.5 Visualization Analysis

To show the recognition accuracy of DTHN in different scenes, several scenes are selected as demonstrations as shown in Fig. 8. The green bounding box indicates the detection results that are higher than 0.5 similarity.

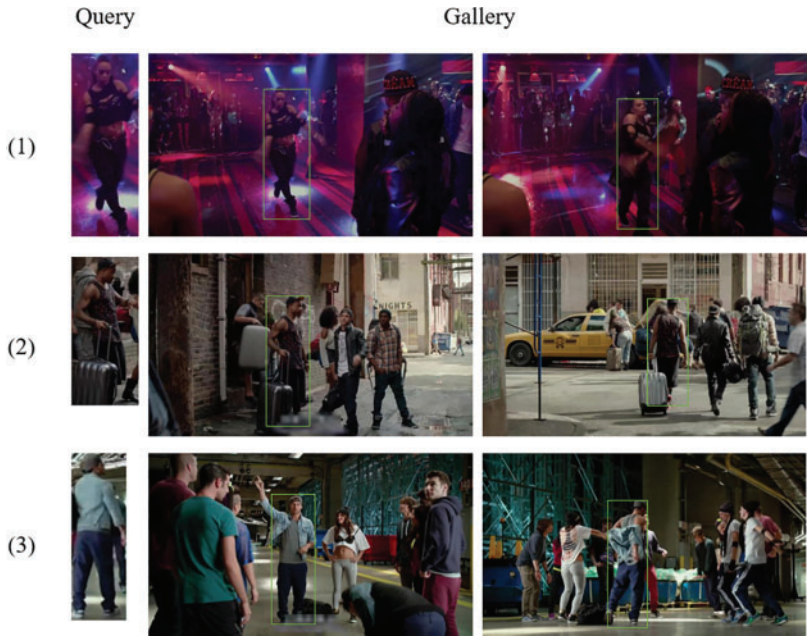

Figure 8: Visualization results of DTHN

Person search is difficult for several reasons, such as camera distance, occlusion, resolution, complex background, and lighting environments. DTHN can extract the features of the target well, thanks to the inclusion of DTH structure. The visualization demonstrates the model's ability to make sound judgments despite a variety of difficult situations, proving the model's effectiveness.

The network takes the query picture as the target and search the person in the gallery. In case (1), the target is a dancing girl on the dance floor. Despite the dim lighting and the fact that dance movements may make the target difficult to recognize, the model is still able to find the target among the many dancers in the scene. In case (2), the target is a young man with a suitcase which covered his lower half body. Despite the lack of information about the lower half, the model can still target in multi-crowd scenarios based on existing information, even with the target's back toward the camera. In case (3), the target is a male with his back to the camera. In the absence of front side information, the model does a good job of identifying the target based on other information such as clothing. In the same back scene with target undressing, the model is still able to correctly recognize the target.

5 Conclusion and Outlook

After noticing the challenges of occlusion and efficiency in end-to-end person search, we propose a DTHN to address the problems. We use two Transformer heads to deal with box detection and re-ID tasks separately, handling high-quality bounding box feature extraction and high-quality re-ID feature extraction. DTHN outperforms existing methods in the CUHK-SYSU dataset and achieves competitive results in the PRW dataset, which demonstrates the method's superior structural design and effectiveness.

Although our method is slightly slower than traditional CNN methods due to the scale dot production used by the attention mechanism in the Transformer, which consumes more computational resources. However, thanks to the small size of the Transformer, we have cut down the number of parameters compared to traditional CNNs, which gives us hope for deployment on terminal devices. Despite the good results, we believe that there is still room for improvement in our approach, either in terms of better and more convenient attention computation methods or in terms of adaptive attention mechanisms. Eventually, we may be able to create a pure Transformer model, using different attention heads on a single Transformer to accomplish different tasks. This is the main focus of our team afterward. We believe that the deployment of person search on terminal devices is just around the corner.

Acknowledgement: Thank you to laboratory colleagues for their support of this paper.

Funding Statement: This research is supported by the Natural Science Foundation of Shanghai under Grant 21ZR1426500, and the National Natural Science Foundation of China under Grant 61873160.

Author Contributions: The authors confirm their contribution to the paper as follows: study conception and design: Cheng Feng; data collection: Cheng Feng; analysis and interpretation of results: Cheng Feng; draft manuscript preparation: Cheng Feng, Dezhi Han, Chongqing Chen. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available upon request from the corresponding author, Cheng Feng, upon reasonable request.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] D. Chen, S. Zhang, W. Ouyang, J. Yang and Y. Tai, "Person search via a mask-guided two-stream CNN model," in *Computer Vision–ECCV, 2018*, Munich, Germany, pp. 764–781, 2018. https://doi.org/10.1007/978-3-030-01234-2_45

[2] W. Dong, Z. Zhang, C. Song and T. Tan, "Instance guided proposal network for person search," in *2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 2582–2591, 2020. https://doi.org/10.1109/CVPR42600.2020.00266

[3] C. Han, "Re-ID driven localization refinement for person search," in *2019 IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Seoul, Korea (South), pp. 9813–9822, 2020. https://doi.org/10.1109/ICCV.2019.00991

[4] X. Lan, X. Zhu and S. Gong, "Person search by multi-scale matching," in *Computer Vision–ECCV 2018*, Munich, Germany, pp. 553–569, 2018. https://doi.org/10.1007/978-3-030-01246-5_33

[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 Conf. on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 11949–11958, 2020. https://doi.org/10.1109/CVPR42600.2020.01197

[6] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang *et al.,* "Person re-identification in the wild," in *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 3346–3355, 2017. https://doi.org/10.1109/CVPR.2017.357

[7] X. Zhou, Y. Zhong, Z. Cheng, F. Liang and L. Ma, "Adaptive sparse pairwise loss for object re-identification," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Vancouver, Canada, pp. 19691–19701, 2023.

[8] Y. Yao, T. Gedeon and L. Zheng, "Large-scale training data search for object re-identification," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Vancouver, Canada, pp. 15568–15578, 2023.

[9] J. Feng, A. Wu and W. S. Zheng, "Shape-erased feature learning for visible-infrared person re-identification," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Vancouver, Canada, pp. 22752–22761, 2023.

[10] G. Zhang, Y. Zhang, T. Zhang, B. Li and S. Pu, "PHA: Patch-wise high-frequency augmentation for transformer-based person re-identification," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Vancouver, Canada, pp. 14133–14142, 2023.

[11] X. Chang, P. Y. Huang, Y. D. Shen, X. Liang, Y. Yang *et al.,* "RCAA: Relational context-aware agents for person search," in *Computer Vision–ECCV 2018*, Munich, Germany, pp. 86–102, 2018. https://doi.org/10.1007/978-3-030-01240-3_6

[12] Z. Li and D. Miao, "Sequential end-to-end network for efficient person search," in *Proc. of the AAAI Conf. on Artificial Intelligence*, Vancouver, British Columbia, Canada, pp. 2011–2019, 2021. https://doi.org/10.1609/aaai.v35i3.16297

[13] T. Xiao, S. Li, B. Wang, L. Lin and X. Wang, "Joint detection and identification feature learning for person search," in *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 3376–3385, 2017. https://doi.org/10.1109/CVPR.2017.360

[14] Y. Yan, "Anchor-free person search," in *2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 7686–7695, 2021. https://doi.org/10.1109/CVPR46437.2021.00760

[15] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu *et al.,* "Learning context graph for person search," in *2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 2153–2162, 2019. https://doi.org/10.1109/CVPR.2019.00226

[16] Y. Zhong, X. Wang and S. Zhang, "Robust partial matching for person search in the wild," in *2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 6826–6834, 2020. https://doi.org/10.1109/CVPR42600.2020.00686

[17] R. Yu, "Cascade transformers for end-to-end person search," in *2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Orleans, LA, USA, pp. 7257–7266, 2022. https://doi.org/10.1109/CVPR52688.2022.00712

[18] 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. https://doi.org/10.1109/TPAMI.2016.2577031

[19] T. Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal loss for dense object detection," in *2017 IEEE Int. Conf. on Computer Vision (ICCV)*, Venice, Italy, pp. 2999–3007, 2017. https://doi.org/10.1109/ICCV.2017.324

[20] Z. Tian, C. Shen, H. Chen and T. He, "FCOS: Fully convolutional one-stage object detection," in *2019 IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Seoul, Korea (South), pp. 9626–9635, 2019. https://doi.org/10.1109/ICCV.2019.00972

[21] A. Vaswani, "Attention is all you need," in *Proc. of the 31st Int. Conf. on Neural Information Processing Systems*, Red Hook, NY, USA, pp. 6000–6010, 2017.

[22] A. Dosovitskiy, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Int. Conf. on Learning Representations*, Vienna, Austria, 2023. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy
[23] Z. Guo and D. Han, "Sparse co-attention visual question answering networks based on thresholds," *Applied Intelligence*, vol. 53, no. 1, pp. 586–600, 2023. https://doi.org/10.1007/s10489-022-03559-4
[24] C. Chen, D. Han and C. C. Chang, "CAAN: Context-aware attention network for visual question answering," *Pattern Recognition*, vol. 132, pp. 108980, 2022. https://doi.org/10.1016/j.patcog.2022.108980
[25] X. Shen, D. Han, Z. Guo, C. Chen, J. Hua *et al.,* "Local self-attention in transformer for visual question answering," *Applied Intelligence*, vol. 53, no. 13, pp. 16706–16723, 2023. https://doi.org/10.1007/s10489-022-04355-w
[26] Z. Guo and D. Han, "Multi-modal co-attention relation networks for visual question answering," *The Visual Computer*, 2022. https://doi.org/10.1007/s00371-022-02695-9
[27] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang *et al.,* "Diverse part discovery: Occluded person re-identification with part-aware transformer," in *2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 2897–2906, 2021. https://doi.org/10.1109/CVPR46437.2021.00292
[28] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770–778, 2016. https://doi.org/10.1109/CVPR.2016.90
[29] J. Xiao, Y. Xie, T. Tillo, K. Huang, Y. Wei *et al.,* "IAN: The individual aggregation network for person search," *Pattern Recognition*, vol. 87, pp. 332–340, 2019. https://doi.org/10.1016/j.patcog.2018.10.028
[30] H. Liu, "Neural person search machines," in *2017 IEEE Int. Conf. on Computer Vision (ICCV)*, Venice, Italy, pp. 493–501, 2017. https://doi.org/10.1109/ICCV.2017.61
[31] B. Munjal, S. Amin, F. Tombari and F. Galasso, "Query-guided end-to-end person search," in *2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 811–820, 2019. https://doi.org/10.1109/CVPR.2019.00090
[32] D. Chen, S. Zhang, W. Ouyang, J. Yang and B. Schiele, "Hierarchical online instance matching for person search," in *Proc. of the AAAI Conf. on Artificial Intelligence*, New York, NY, USA, pp. 10518–10525, 2020. https://doi.org/10.1609/aaai.v34i07.6623
[33] W. Dong, Z. Zhang, C. Song and T. Tan, "Bi-directional interaction network for person search," in *2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 2836–2845, 2020. https://doi.org/10.1109/CVPR42600.2020.00291
[34] D. Chen, S. Zhang, J. Yang and B. Schiele, "Norm-aware embedding for efficient person search," in *2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 12612–12621, 2020. https://doi.org/10.1109/CVPR42600.2020.01263
[35] C. Han, Z. Zheng, C. Gao, N. Sang and Y. Yang, "Decoupled and memory-reinforced networks: Towards effective feature learning for one-step person search," in *Proc. of the AAAI Conf. on Artificial Intelligence*, Vancouver, British Columbia, Canada, pp. 1505–1512, 2021. https://doi.org/10.1609/aaai.v35i2.16241
[36] H. Kim, S. Joung, I. J. Kim and K. Sohn, "Prototype-guided saliency feature learning for person search," in *2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 4863–4872, 2021. https://doi.org/10.1109/CVPR46437.2021.00483
[37] B. J. Han, K. Ko and J. Y. Sim, "End-to-end trainable trident person search network using adaptive gradient propagation," in *2021 IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 905–913, 2021. https://doi.org/10.1109/ICCV48922.2021.00096
[38] Y. Yan, J. Li, J. Qin, P. Zheng, S. Liao *et al.,* "Efficient person search: An anchor-free approach," *International Journal of Computer Vision*, vol. 131, pp. 1642–1661, 2023. https://doi.org/10.1007/s11263-023-01772-3
[39] M. Fiaz, H. Cholakkal, R. M. Anwer and F. S. Khan, "SAT: Scale-augmented transformer for person search," in *Proc. of the IEEE/CVF Winter Conf. on Applications of Computer Vision*, Waikoloa, HI, USA, pp. 4820–4829, 2023.