SAT: Scale-Augmented Transformer for Person Search

Mustansar Fiaz, Hisham Cholakkal, Rao Muhammad Anwer, Fahad Shahbaz Khan
Department of computer Vision,
Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE.
(mustansar.fiaz, hisham.cholakkal, rao.anwer, fahad.khan)@mbzuai.ac.ae

Abstract

Person search is a challenging computer vision problem where the objective is to simultaneously detect and re-identify a target person from the gallery of whole scene images captured from multiple cameras. Here, the challenges related to underlying detection and re-identification tasks need to be addressed along with a joint optimization of these two tasks. In this paper, we propose a three-stage cascaded Scale-Augmented Transformer (SAT) person search framework. In the three-stage design of our SAT framework, the first stage performs person detection whereas the last two stages performs both detection and re-identification. Considering the contradictory nature of detection and re-identification, in the last two stages, we introduce separate norm feature embeddings for the two tasks to reconcile the relationship between them in a joint person search model. Our SAT framework benefits from the attributes of convolutional neural networks and transformers by introducing a convolutional encoder and a scale modulator within each stage. Here, the convolutional encoder increases the generalization ability of the model whereas the scale modulator performs context aggregation at different granularity levels to aid in handling pose/scale variations within a region of interest. To further improve the performance during occlusion, we apply shifting augmentation operations at each granularity level within the scale modulator. Experimental results on challenging CUHK-SYSU [35] and PRW [47] datasets demonstrate the favorable performance of our method compared to state-of-the-art methods. Our source code and trained models are available at this https URL.

1. Introduction

Person search [35] is a promising and challenging research area which localizes and discriminates the identity of a particular query person in a gallery of real-world scene frames. A person search problem can be identified as a unified system where two isolated objectives (i.e., detection [3, 30, 32, 44] and re-identification [41, 15, 25, 28]) are performed together. The person search is highly complicated problem due to person detection and re-identification challenges along-with the joint optimization of these sub-tasks. In real-world scenarios, a person search algorithm must locate and identify the target person from complex scenarios such as pose/view variations, appearance variations, scale variations, occlusions or background clutters.

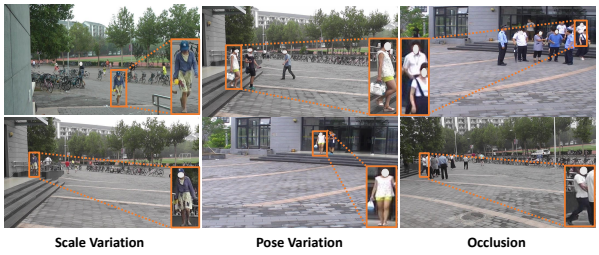

Figure 1. The major three challenges in person search problem such as scale variation, pose variation, and occlusion.

Various efforts devoted in person search problem can be roughly categorized into two-step [6, 12, 24] and one-step [27, 29, 7] approaches. In two-step approaches, the detection and re-identification (ReID) tasks are decoupled and performed sequentially. The pedestrians are first localized with off-the-shelf detectors and later fed into an ReID network to identify the pedestrians from the cropped person patches. Despite their promising results, these approaches lack in computational efficiency. In contrast, one-step approaches unify person detection and re-identifications using a single network. Such approaches [7, 26] extend two-stage detectors such as Faster-RCNN by employing additional ReID loss for person identity discrimination. Nevertheless, aforementioned approaches are still lacking in three major issues discussed below:

- The person search problem mainly strives for the conflict between the person detection and person re-identification [7, 6]. The objective of detection is to classify the people from the background using shared feature embedding, whereas ReID discriminates the identities of the people. Chen et al. [7] introduced Norm-Aware Embedding (NAE) to decompose the

feature embedding, in polar coordinate system, to radial norm and angle for detection and ReID tasks respectively. Later on, this strategy was utilized in various works [7, 26, 42, 17]. However, the parameters for NAE are still shared between the detection and ReID sub-tasks resulting in sub-optimal solution.

- A person can undergo scale and pose variations as shown in Fig. 1 in a challenging scene, that increases the complexity of person identity. To handle these challenges, various attempts using either feature pyramids or deformable convolutions [38, 7, 46] have been made. Howbeit, feature fusion may add background noise leading to inferior ReID performance.

- Moreover, the appearance deformations and occlusion, as shown in Fig. 1, may deteriorate the region of interest (RoI) features quality resulting in imprecise identity discrimination. Although, most of the previous work achieved improved accuracy, they are at the disposal of failure due to the holistic appearance representation of people in one-step [33, 26] or two-step approaches [38].

To overcome above-mentioned challenges, we propose a hybrid context aggregator to fuse the merits of CNNs and ViT [14] into a cascaded end-to-end person search method. We utilize coarse-to-fine strategy just like cascaded-RCNN [1] to improve the quality of detection and re-identification at different stages. In the first stage, we perform person detection and generalization across the people without identity discrimination. Whereas, in later stages, we refine both detection and ReID embeddings based on previous stage regression estimations. To be specific, to tackle the first challenge, we reduce the contradictory objective between the person detection and person ReID by explicitly decoupling the NAE feature representations for both sub-tasks. This decoupled NAE feature representation reduces the dependency upon each other and improves the detection as well as identity similarity confidence. Secondly, we propose a Scale-Augmented Transformer (SAT) network at each stage to deal with different scale/pose variations and occlusions. The SAT network passes the base features into a convolutional encoder to increase the generalization [36] and than to a transformer to capture global level instance information. Specifically, the output of convolutional encoder is split into two parts. We apply depth-wise convolutions at different granularity levels on half set of features and finally fuse the modulated features into the remaining half set of features. Thirdly, to cope the appearance deformations and occlusions, we split the features and apply different augmentations at misaligned tokens via shifting operation across each sub-feature. Later on, these sub features are fused after mixing via depth-wise convolution. Experimental study over PRW [47] and CUHK-SYSU [35] datasets shows the superiority of the proposed PS system.

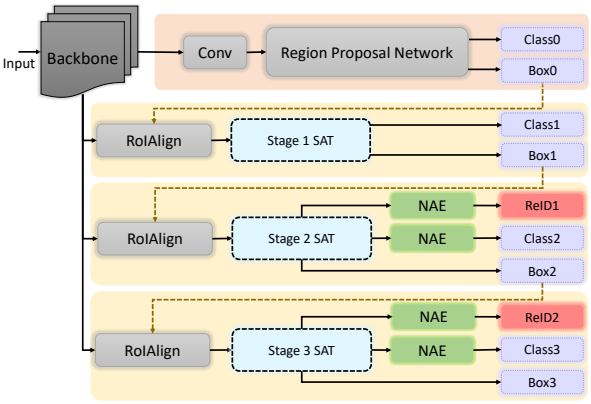

Figure 2. Our proposed cascaded person search framework. The input is passed to backbone network to produce stem features. These features are passed to RPN using Conv layers. The backbone features along with bounding boxes are also forwarded to different stages to obtain RoIAlign pooled features. All stages take bounding boxes from previous stage bounding box estimations except the first stage, which takes bounding boxes from RPN. The RoIAlign pooled features are fed to the proposed SAT network at stage 1, stage 2 and stage 3. All three stages are optimized with class and box heads, while stage 2 and stage 3 are optimized with an additional ReID head. Moreover, detection and ReID tasks are decoupled for stage 2 and stage 3 by introducing a separate NAE for both tasks.

1.1. Contribution:

- We explicitly decouple the norm-aware representations between detection and ReID, which leads to more detection confidence and more identity similarity.

- We propose a context aggregator block to leverage the advantages from both the CNNs and transformers for person search.

- To handle scale/pose variations, we propose a scale-aware network that implicitly aggregates the scale information within each RoI from different scales.

- In order to deal with the occlusion/deformation within an RoI, we employ different augmentations at misaligned tokens via shifting and mixing operations.

- Extensive experiments on two datasets exhibit the advantages of the proposed method compared to state-of-the-art approaches.

2. Related Work

2.1. Person Search

Person Re-identification have shown immense achievements in the field [40, 37, 34], where the query person is

matched with the gallery of cropped person images. However, there exists a research gap to apply ReID problem in the real-world applications. Therefore, person search is introduced with aim to localize and identify the query persons from the set of full resolution images of a scenario [35]. The previous works on this can be broadly classified as two-step and one-step models. In two-step models, the target person detection and re-identifications are performed independently in a sequential manner [12, 18, 24, 33]. For example, Wang et al. [33] introduced TCTS method to deal with the inconsistent relationship between the detection and ReID. Lan et al. [24] proposed a multi-scale feature pyramid for person re-identification.

In contrast, one-step models do detection and ReID in a joint framework, which make them more efficient and effective [7, 38]. Ever since the introduction of Faster RCNN [32], numerous one-step person search have been proposed [7, 11, 26, 17, 5, 29]. Chen et al. [7] used norm-aware embeddings to detach the person embedding for detection and reID. Munjal et al. [29] used a query-guided Siamese squeeze and excitation block to exploit the relationship between person and gallery images. Dong et al. [11] proposed a BINet that takes both entire and cropped images into a Siamese network for better person feature representation learning. Yan et al. [38] introduced an anchor free person search framework. Recently, Li and Miao [26] proposed a SeqNet which employs two faster RCNN network in a sequential manner for detection and ReID. Although, these approaches provide satisfactory results, they strive from the conflicting objective between the detection and re-identification and share the same norm feature embeddings. In contradiction, we introduce a separate norm feature embeddings for the both subtasks to further release the conflicting embeddings. Moreover, we utilize a cascaded approach to refine RoI pooled features at multiple stages.

2.2. Transformer based Approaches

Since the advent of ViT model [14] for image recognition task, it is being used in several computer vision applications including person re-identification [34, 25, 43]. Wang et al. [34] proposed neighbor transformer by exploiting the neighbouring features to obtain robust representation for person re-identification. Zhang et al. [43] used a transformer based feature calibration approach for person re-identification by using low-level feature information as a global priors. Li et al. [25] proposed a part discovery technique by using part-aware transformer to deal with occlusion for person ReID. Recently, PSTR [2] and COAT [42] introduced transformers in the person search pipeline. PSTR is based on DETR [4] framework, which utilizes encoder-decoder architecture for detection and decoder for re-identification. On the other hand, COAT [42] is based on cascaded RCNN [1] to learn the discriminative coarse-to-fine representations at multiple stages. It uses explicit-multiscale convolution transformers to deal with scale variations at each stage. On the contrary, we propose a context aggregator for person search to benefit from the intrinsic properties of CNNs and transformers. We propose an implicit transformer-based architecture that take cares of scale variations at each stage. Furthermore, in contrast to COAT, we use different augmentation techniques at different misaligned tokens via shifting and mixing operations to synthetically alleviate the occlusions.

3. Method

3.1. Overall Architecture

The overall architecture of the proposed person search framework having three stages is shown in Fig. 2. Since the person search has conflicting objectives between detection and ReID, we induce separate norm-aware feature representations for both sub-tasks. Besides, our design introduces a hybrid context aggregator at each stage to benefit from the inherit characteristics of CNNs and transformers. Considering that the model performance may deteriorate due to scale/pose variations as well as occlusions, we propose scale-augmented transformer to refine the detection and ReID successively at multiple stages.

We use ResNet-50 [20] backbone network to generate the 1024-stem features maps and are passed to each stage. In first stage, we get the proposals from Region Proposal Network (RPN) [32]. In addition, the first stage is optimized by using detection and regression heads, while the last two stages are optimized by employing the detection, regression, and re-identifications heads based on regression estimations at the previous stages.

3.2. Decoupled Detection and ReID Embedding

In the faster RCNN based person search framework, the objective of detection is to perform inter-class discrimination between the target object from background, while ReID is responsible for intra-class discrimination to identify the particular person. This approach suffers from conflicting objectives between detection and ReID using the same backbone network. Therefore, stepping forward to mitigate the aforementioned conflict, we explicitly decouple the Norm-Aware Embedding (NAE) representations for ReID and detection. Specifically, we introduce an independent NAE representation for detection as well as ReID to reconcile the relationship between the detection and ReID.

3.3. Scale-Augmented Transformer

Since a person may undergo several scale and pose variations in the scene, it is desirable to learn these variations without any supervision. To this end, we propose a hybrid context aggregator for person search called Scale-

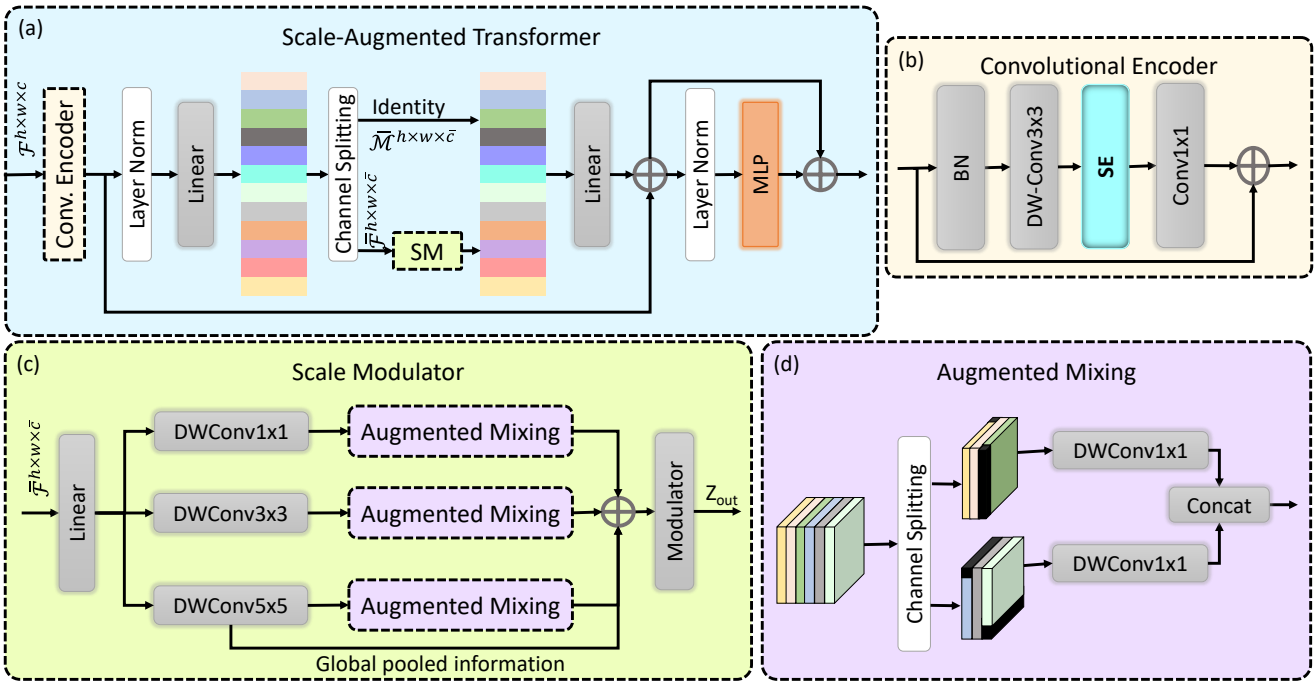

Figure 3. (a) Overall architecture of proposed Scale-Augmented Transformer (SAT) network. It comprises of convolutional encoder, two norm layers , two linear layers, scale modulator (SM), and a MLP block. Our design focuses to leverage from both the CNNs and the transformer. The input RoI pooled features are fed to convolutional encoder to increase the generalization and discriminative ability. The output of convolutional encoder is passed to norm layer followed by a linear layer and is split channel-wise. One half set of features $\bar{\mathcal{F}} \in \mathcal{R}^{h \times w \times \bar{c}}$ is fed to a scale modulator (c) to learn the local regions of an object at various granularity levels. The scale modulator input realized with depth-wise convolutions and forward to augmented mixing (d). Here, we split the features and apply different augmentations at misaligned tokens via shifting operation across each sub-feature. Later on, these sub features are fused after mixing via depth-wise convolution. The output of augmented mixing at different granularity levels is gathered along with global information are fused, and passed to a modulator. Finally, the output of scale modulator is fused with other half set of features $\bar{\mathcal{M}} \in \mathcal{R}^{h \times w \times \bar{c}}$ and passed to a linear layer. The output is added with the output of convolutional encoder and fed to a norm layer followed by a mlp block.

Augmented Transformer (SAT), which is composed of a convolutional encoder block, two linear layers, two normalization layers, Scale Modulator (SM) block, and MLP block. The proposed hybrid SAT network strives to explicitly combine the strength of CNNs to capture local features as well as transformer to encapsulate long range dependencies. Motivated by [22], we include a convolutional encoder block prior to the proposed SAT network which improves the generalization and discriminative ability [36] of the ReID. The block diagram of convolutional encoder is shown in Figure 3-(b). Additionally, it reduces the requirement of conventional position embedding layers known as tokenization in ViTs [13] due to intrinsic properties of depth-wise convolutions, that can be considered as conditional positional embedding [9].

To this end, the RoI input features are empowered using two convolutional layers and a normalization layer between them to obtain desirable dimensional features $\mathcal{F} \in \mathcal{R}^{h \times w \times c}$, which are fed to convolutional encoder block following a norm layer and a linear layer. The output of linear layer is split channel-wise into two halves. One half set of feature $\bar{\mathcal{F}} \in \mathcal{R}^{h \times w \times \bar{c}}$ is passed to the scale modulator to learn the scale of person at different granularity levels. The scale modulator empirically acts as a persistent network to encode the pose and scale variations in an explicit manner among the local regions of an object at various scales. The output features of scale modulator along with the identical other half $\bar{\mathcal{M}} \in \mathcal{R}^{h \times w \times \bar{c}}$ are concatenated and forwarded to a linear layer. This output is fused with the output of the convolutional encoder block using skip connection. Afterwards, a channel-wise mixing is employed using a norm layer and MLP block as shown in Figure 3-(a). Finally, the output of SAT network is linearly transformed into the expected dimension. Note that, there is a residual connection outside the SAT network. Eventually, after Global Average Pooling (GAP), the features are forwarded into individual heads i.e., regressor, NAE detection, and NAE ReID.

3.4. Scale Modulation

To learn/encode the scale and pose variations of the query person within a RoI, we introduce a Scale Modulator (SM) as shown in Figure 3-(c). Yu et al. [42] explicitly first utilizes convolutional layer with different kernels to obtain features at different scales, and then passes to a transformer. Although this approach returns satisfactory results, each feature channel does not tackle the scale efficiently due to diverse variations in the gallery images, which may yield sub-optimal solution. In contrast, we studiously propose an implicit scale modulator by leveraging the benefit of modulation operation. Another advantage of the proposed method is that channel mixing is required only once for all scales instead of applying it for each scale, which reduces the computational complexity of the model as well.

The features $\bar{\mathcal{F}} \in \mathcal{R}^{h \times w \times \bar{c}}$ are realized with a linear layer and passed to depth-wise convolutions with three different kernels revealing the features at three different scales. These features are passed to an augmented mixing (discussed in next section 3.5). The augmented mixed features at different scales as well as global pooled information are fused and send to a modulator forced by a convolutional layer.

3.5. Augmented Mixing

To deal with appearance deformations in an RoI, He et al. [21] shuffles the person parts which might contains different parts of multiple people. On the other hand, Yu et al. [42], exchanges the partial tokens of the people in a mini batch which may learn the inaccurate partial information of different instances. In contrast, we introduce a specialized augmentation mixing technique to learn the robust representation against appearance deformations/occlusions and misalignments within an RoI.

We split the input features into channel-wise and perform shift augmentation. Precisely, we first pad a zero vector in a particular direction (for example, on the left side), perform a single shift operation in that direction and remove the vector in opposite direction (i.e., from the right side) to obtain the same size feature maps. Similarly, we perform such shifted augmentation across all four directions. These augmented features are mixed using depth-wise convolutions and fused using a concatenation operation as shown in Fig.3-(d). Note that, there is a residual connection across the augmented mixing, yielding robust representation against partial occlusions.

4. Experiments

To validate the effectiveness of the proposed method, we evaluate our approach on two well known datasets PRW [47] and CUHK-SYSU [35]. The following section discusses datasets, metrices, and experimental details. Furthermore, performance comparisons with the state-of-the-art (SOTA) are presented on these datasets. Finally, an ablation study is performed to endorse the potency of the proposed algorithm.

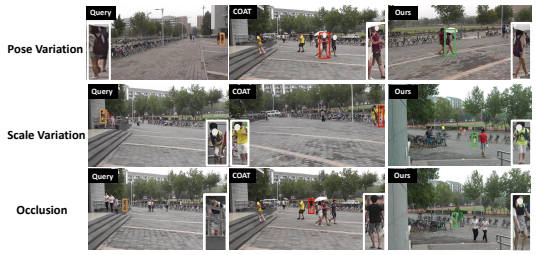

Figure 4. Qualitative comparison between the COAT [42] and ours method in three different challenging scenarios. For all cases, our method shows correct top-1 matching results. The orange, red, and green colors indicate the query, failure, and correct respectively.

4.1. Datasets and Settings

4.1.1 CUHK-SYSU

CUHK-SYSU [35] is a large-scale person search dataset which contains heterogeneous real-world challenges, such as illumination variations, scale variations, pose variations, resolution, occlusion, and diverse backgrounds. There are a total of 18,184 images where 96,143 are the annotated pedestrians with 8,432 different identities. The dataset adopts the standard train and test sets. The train set has 5,532 identities and 6,978 frames, whereas test set contains 2,900 query people and 6,978 frames. Moreover, this dataset provides a range of gallery sizes from 50 to 4,000 to report the scalability of the model. We report results on standard gallery size of 100 unless it is specified otherwise.

4.1.2 PRW

The PRW dataset [47] is acquired using six static cameras in a university. It contains a total of 11,816 images with 43,110 manually annotated bounding boxes, where 34,304 are annotated as people with 932 identities and remaining boxes are marked as unknown identities. The dataset is split into train and test sets. The train set contains 5,704 images with 482 identities and test set has 2,057 query persons which are searched in a gallery of 6,112 frames. Hence, the gallery set is significantly larger than the CUHK-SYSU dataset.

4.1.3 Evaluation Protocols

We follow conventional protocols to evaluate the person search including the mean average precision (mAP) and top-1. To compute the detection performance, we also used average precision (AP) and recall.

Figure 5. Qualitative analysis on CUHK-SYSU [35] (top 2 rows) and PRW [47] (bottom row) datasets. We show the top two matching results for different query person. Our method correctly detects and identifies the query persons under different indoor and outdoor scenarios.

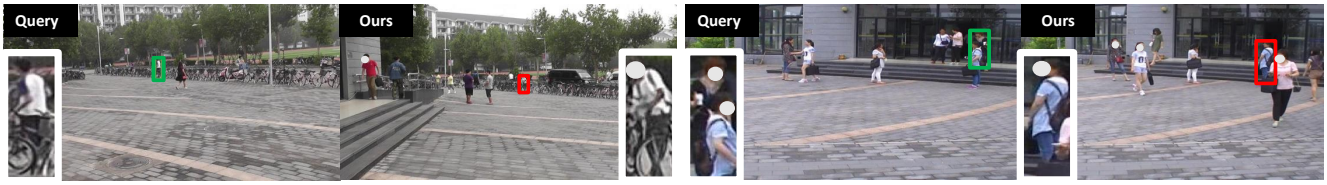

Figure 6. Failure cases on PRW [47] dataset. We show that our method incorrectly identifies query person under heavy occlusions.

4.1.4 Implementation Details

We use ResNet-50 [20] backbone network trained over ImageNet dataset [10]. The proposed method is implemented in python using PyTorch [31] library. We adopt COAT [42] as our baseline network and used three stage cascaded framework and extract 128 detection proposals at every stage. Similar to faster-RCNN [32] based approaches, we set width w and height h to 14 [7, 26]. We set IoU threshold as 0.5, 0.6, and 0.7 for detection in three stages respectively. Furthermore, similar to COAT [42], we include an additional cross-entropy loss for identity supervision at second and third stages only. The network is trained using SGD optimizer with momentum 0.9 for 12 epochs. The initial learning rate is set to 0.003 with warm up at first epoch and is decreased at 10th epoch. Moreover, during inference, we use NMS with thresholds 0.4, 0.4, and 0.5 for three consecutive stages respectively to eliminate the redundant bounding boxes.

4.2. Comparison with state-of-the-art methods

In this section, we compare our method with two-step and one-step state-of-the-art methods in Table 2.

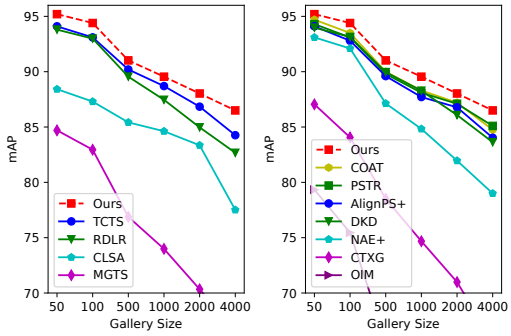

Figure 7. Performance comparison on CUHK-SYSU with varying gallery sizes. The dash line indicating consistent better performance compared to both two-step and one-step approaches represents our method.

4.2.1 Comparison on CUHK-SYSU dataset

The performance on CUHK-SYSU dataset is compared using gallery size of 100. Among two-step methods, TCTS [33] performs best with 93.9% mAP and 95.1% top-1 scores. On the other hand, AlignPS+ [38] and COAT [42] performs better with 94.0% and 94.2% mAP scores among one-step methods. In contrast, our method outperforms in terms of both mAP 94.4% and 94.8% top-1. Compared to

the recently introduced PSTR [2] with strong object detector achieves mAP of 93.5%, our method exceeds in terms of mAP and demonstrates comparable result using top-1.

We further perform experiments on CUHK-SYSU dataset with varying gallery set from 50 to 4000, which increases the gallery complexity due to more complex scenarios such as scale/pose variations and occlusions. Figure 7 demonstrates that our method consistently achieves better accuracy compared to existing two-step as well as one step methods on different gallery sizes. Although transformer based COAT and PSTR, and CNNs based AlignPS+ and DKD exhibit similar performance, our approach composed of hybrid context aggregator shows consistent performance improvement over varying gallery sets. This is demonstrating the ability of the proposed method to tackle scale variations and occlusions efficiently in the large gallery sets.

4.2.2 Comparison on PRW dataset

Compared to CUHK-SYSU dataset, PRW dataset has large gallery size with less available training data. Therefore, PRW dataset is more challenging. Among existing two-step methods, our method exceeds the top performing MGN+OR [39] and TCTS [33] and achieves 54.5% mAP and 87.5% top-1. Among one-step SOTA, AGWF [17] with part classification, SeqNet [26] with two-stage refinement, and COAT [42] with three stage refinement, our method performs better with 54.5% mAP. Our method achieves a significant gain of 5.0% compared to recently introduced PSTR [2] with stronger DETR object detector [4]. In terms of top-1, our method achieves 87.5% which is comparable with PSTR [2] and AGWF [17].

4.2.3 Qualitative performance

We first compare our method qualitatively with COAT [42] over PRW dataset. Figure 4 shows that our method detects and identifies the query person successfully in various challenging scenes. Our method shows performance improvement due to decoupled NAE, hybrid context aggregator and implicit scale-augmented transformer to handle scale/pose variations. We present qualitative results over CUHK-SYSU and PRW datasets in Figure 5. This represents that our method can correctly localize and recognize query people under challenging scenarios. We also show the failure cases in Figure 6 where the query person is heavily occluded.

4.2.4 Efficiency Comparison

Here, we evaluate the efficiency for different person search methods. It is difficult to perform a fair comparison where the methods are evaluated over different GPUs. Therefore, we show the Tera-Floating Point Operation per second (TFLOPs) for each GPU. To keep consistent with other methods, the input images are resized to 900×1500. From Table 1, we see that our method is about 2 times faster than the realtime MGTS and QEEPS methods. Although our method demonstrates slightly slower speed compared to newly introduced PSTR and COAT methods, it achieves an absolute gain of 5.0% and 1.3% in mAP. It also reveals the potential for realworld applications.

Table 1. Speed vs accuracy comparison for person search methods over PRW dataset. Time is in milliseconds.

Methods	GPU (TFLOPs)	mAP	Time (ms)
MGTS [6]	K800 (4.1)	32.6	1269
QEEPS [29]	P6000 (12.6)	37.1	300
DKD [45]	1050Ti (11.3)	50.5	124
NAE [7]	V100 (14.1)	43.3	83
NAE+ [7]	V100 (14.1)	44.0	98
SeqNet [26]	V100 (14.1)	46.7	86
AlignPS[38]	V100 (14.1)	45.9	61
PSTR [2]	V100 (14.1)	49.5	56
COAT [42]	V100 (14.1)	53.3	90
Ours	V100 (14.1)	54.5	105

Table 2. State-of-the-art comparison on CUHK and PRW test sets using mAP and top-1 accuracy. Our SAT performs better as compared to two-step and one-step state-of-the-art methods.

	Method	CUHK-SYSU		PRW	
		mAP	top-1	mAP	top-1
Two-step	CLSA [24]	87.2	88.5	38.7	65.0
	IGPN [12]	90.3	91.4	42.9	70.2
	RDLR [18]	93.0	94.2	42.9	70.2
	MGTS [6]	83.0	83.7	32.6	72.1
	MGN+OR [39]	93.2	93.8	52.3	71.5
	TCTS [33]	93.9	95.1	46.8	87.5
End-to-end	OIM [35]	75.5	78.7	21.3	49.9
	QEEPS [29]	88.9	89.1	37.1	76.7
	HOIM [5]	89.7	90.8	39.8	80.4
	BINet [11]	90.0	90.7	45.3	81.7
	PGSFL [23]	92.3	94.7	44.2	85.2
	DKD [45]	93.1	94.2	50.5	87.1
	APNet [48]	88.9	89.3	41.2	81.4
	DMRN [19]	93.2	94.2	46.9	83.3
	CAUCPS [16]	81.1	83.2	41.7	86.0
	ACCE [8]	93.9	94.7	46.2	86.1
	AlignPS [38]	93.1	93.4	45.9	81.9
	AlignPS + [38]	94.0	94.5	46.1	85.8
	NAE [7]	91.5	92.4	43.3	80.9
	NAE+ [7]	92.1	94.7	44.0	81.1
	AGWF [17]	93.3	94.2	53.3	87.7
	SeqNet [26]	93.8	94.6	46.7	83.4
	PSTR [2]	93.5	95.0	49.5	87.8
	COAT [42]	94.2	94.7	53.3	87.4
	AlignPS [38] + CBGM [26]	93.6	94.2	46.8	85.8
	AlignPS + [38] + CBGM [26]	94.2	94.3	46.9	85.7
	SeqNet [26] + CBGM [26]	94.8	95.7	47.6	87.6
	PSTR [2] + CBGM [26]	-	-	50.1	89.2
	COAT[42] + CBGM [26]	94.8	95.2	54.0	89.1
	Ours (SAT)	94.4	94.8	54.5	87.5
	Ours (SAT) + CBGM [26]	95.3	96.0	55.0	89.2

4.3. Ablation Study

We present an extensive ablation study on PRW dataset to validate the effectiveness of our approach. Table 3, shows

Table 3. Ablation study over the PRW dataset by gradually adding our novel contributions to the baseline. While adding our SAT network to each stage without convolutional embedding, it increases the mAP but reduces the top-1. Adding convolutional encoder into SAT network benefits from inherit properties of CNNs as well as transformer results in optimal solution.

Method	ReID		Detection	
	mAP	top-1	Recall	AP
Baseline	50.96	85.56	95.51	92.62
Baseline + decoupled NAE	51.80	85.27	93.23	93.39
Baseline + decoupled NAE + scale modulator	53.15	86.19	95.46	93.09
Baseline + decoupled NAE + scale modulator + augmented mixing	53.84	85.51	95.38	93.02
SAT	54.45	87.52	95.50	93.17

Table 4. Comparison of different variant of our SAT over PRW dataset. Introducing SAT network at each stage results in best performance.

Stage1	Stage2	Stage3	mAP	Top-1
Res5	Res5	Res5	51.80	85.27
Conv. Encoder	Conv. Encoder	Conv. Encoder	53.37	86.44
Conv. Encoder	Conv. Encoder	SAT	53.63	86.56
Conv. Encoder	SAT	SAT	54.03	86.78
SAT	SAT	SAT	54.45	87.52

the performance of our incremental contributions to the baseline. We replace proposed SAT network with the res5 block, to make our baseline, as in [26, 7] for each stage. Our baseline provides mAP of 50.96% and top-1 85.56%. As discussed earlier, first we decouple the detection and ReID NAE representation to release the conflict between them, it adds ReID score by 0.84% in terms of mAP and detection score by 0.77% in terms of AP to the baseline. After that, we include our scale modulation network (excluding augmented mixing), which leverages the benefits of scale modulation at different granularity levels. This increases the ReID performance to 53.15% mAP and 86.19% top-1. Subsequently, we introduced our augmented mixing within scale modulation leads to an over all mAP of 53.84% but it reduces the top-1 to 85.51%. Finally, to further complement, we introduce a convolutional encoder to the scale modulator. This results in an increased performance using ReID mAP by 54.45% and top-1 by 87.52%.

Stage-wise comparison: We further evaluate the effectiveness of our contributions at different stages on PRW dataset as shown in Table 4. First, we replace the res5 with proposed convolutional encoder which leads to improved performance by 53.37% mAP and 86.44% top-1. Later, we replace the convolutional encoder with our proposed SAT at multiple stages. We observe that including SAT network in all stages results in optimum solution with 54.45% mAP and 87.52% top-1.

Analysis of Context Aggregator We validate the effectiveness of the proposed hybrid context aggregator on PRW dataset. In table 5, first row indicates the baseline approach with the decoupled NAE, where we place the res5 at all three stages. Replacing res5 with proposed convolutional encoder leads to improved performance by 53.37% mAP and 86.44% top-1. Later, we replace the res5 with the proposed scale-augmented transformer which results in 53.84% mAP and 85.51% top-1. Finally, we combine the convolutional encoder with scale-augmented transformer to form a hybrid context aggregator, yields best 54.45% mAP and 87.52% top-1. This indicates the hybrid context aggregator benefits from both convolutional encoder and transformer which improves the performance.

Table 5. Comparison of different context aggregators over PRW dataset.

Res5	Conv. Encoder	Transformer	mAP	Top-1
✓	-	-	51.80	85.27
-	✓	-	53.37	86.44
-	-	✓	53.84	85.51
-	✓	✓	54.45	87.52

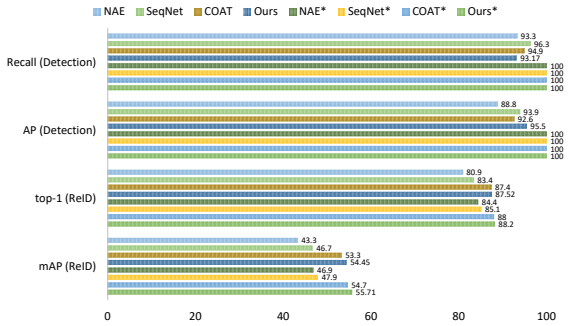

Figure 8. Person search and detection scores on PRW dataset with and without ground-truth detection boxes. The $*$ indicates the results using ground-truth boxes.

Relation between detection and ReID: In Figure 8, we further verify the potency of our method to deal with detection and ReID objectives. We compared our method using predicted detections from the model as well as target boxes from the ground-truth. Among faster RCNN based approaches such as COAT [42], SeqNet [26], and NAE [7], our method demonstrates consistent performance gain with and without ground-truth boxes.

5. Conclusion

We develop a three stage cascaded person search method called SAT to learn the robust ReID representations in a coarse-to-fine manner. Our method accommodates the contradictory relationship between the detection and re-identification by introducing separate feature embeddings for the two subtasks. Moreover, the pivot to our design jointly benefits from the properties of CNNs and transformer. The proposed SAT network employs convolutional encoder to enhance the generalization ability of the model. It aggregates the features at different granularity levels to deal with different scale variations within RoI as well as apply different augmentations at misaligned tokens via shifting operation to tackle occlusions. Extensive experiments performed on two benchmark datasets demonstrate the merits of our novel contributions and state-of-the-art performance.

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[2] Jiale Cao, Yanwei Pang, Rao Muhammad Anwer, Hisham Cholakkal, Jin Xie, Mubarak Shah, and Fahad Shahbaz Khan. Pstr: End-to-end one-step person search with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9458–9467, 2022.

[3] Jiale Cao, Yanwei Pang, Jin Xie, Fahad Shahbaz Khan, and Ling Shao. From handcrafted to deep features for pedestrian detection: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *Proc. European Conference on Computer Vision*, 2020.

[5] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Bernt Schiele. Hierarchical online instance matching for person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10518–10525, 2020.

[6] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. Person search via a mask-guided two-stream cnn model. *Proc. European Conference on Computer Vision*, 2018.

[7] Di Chen, Shanshan Zhang, Jian Yang, and Bernt Schiele. Norm-aware embedding for efficient person search. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[8] Shihui Chen, Yueqing Zhuang, and Boxun Li. Learning context-aware embedding for person search. *arXiv preprint arXiv:2111.14316*, 2021.

[9] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Conditional positional encodings for vision transformers. *arXiv preprint arXiv:2102.10882*, 2021.

[10] 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*, pages 248–255. Ieee, 2009.

[11] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Bi-directional interaction network for person search. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[12] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Instance guided proposal network for person search. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *Proc. International Conference on Learning Representations*, 2020.

[15] Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1060–1069, 2022.

[16] Byeong-Ju Han, Kuhyeun Ko, and Jae-Young Sim. Context-aware unsupervised clustering for person search. *arXiv preprint arXiv:2110.01341*, 2021.

[17] Byeong-Ju Han, Kuhyeun Ko, and Jae-Young Sim. End-to-end trainable trident person search network using adaptive gradient propagation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 925–933, 2021.

[18] Chuchu Han, Jiacheng Ye, Yunshan Zhong, Xin Tan, Chi Zhang, Changxin Gao, and Nong Sang. Re-id driven localization refinement for person search. *Proc. IEEE International Conference on Computer Vision*, 2019.

[19] Chuchu Han, Zhedong Zheng, Changxin Gao, Nong Sang, and Yi Yang. Decoupled and memory-reinforced networks: Towards effective feature learning for one-step person search. *Proc. AAAI Conference on Artificial Intelligence*, 2021.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[21] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022, 2021.

[22] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[23] Hanjae Kim, Sunghun Joung, Ig-Jae Kim, and Kwanghoon Sohn. Prototype-guided saliency feature learning for person search. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[24] Xu Lan, Xiatian Zhu, and Shaogang Gong. Person search by multi-scale matching. *Proc. European Conference on Computer Vision*, 2018.

[25] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2898–2907, 2021.

[26] Zhengjia Li and Duoqian Miao. Sequential end-to-end network for efficient person search. *Proc. AAAI Conference on Artificial Intelligence*, 2021.

[27] Hao Liu, Jiashi Feng, Zequn Jie, Jayashree Karlekar, Bo Zhao, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. Neural person search machines. *Proc. IEEE International Conference on Computer Vision*, 2017.

[28] 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.

[29] Bharti Munjal, Sikandar Amin, Federico Tombari, and Fabio Galasso. Query-guided end-to-end person search. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[30] Yanwei Pang, Jin Xie, Muhammad Haris Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Ling Shao. Mask-guided attention network for occluded pedestrian detection. *Proc. IEEE International Conference on Computer Vision*, 2019.

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Proc. Advances in Neural Information Processing Systems*, 2015.

[33] Cheng Wang, Bingpeng Ma, Hong Chang, Shiguang Shan, and Xilin Chen. Tcts: A task-consistent two-stage framework for person search. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[34] Haochen Wang, Jiayi Shen, Yongtuo Liu, Yan Gao, and Efstratios Gavves. Nformer: Robust person re-identification with neighbor transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7297–7307, 2022.

[35] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[36] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *Advances in Neural Information Processing Systems*, 34:30392–30400, 2021.

[37] Shiyu Xuan and Shiliang Zhang. Intra-inter camera similarity for unsupervised person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11926–11935, 2021.

[38] Yichao Yan, Jingpeng Li, Jie Qin, Song Bai, Shengcai Liao, Li Liu, Fan Zhu, and Ling Shao. Anchor-free person search. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[39] Hantao Yao and Changsheng Xu. Joint person objectness and repulsion for person search. *IEEE Transactions on Image Processing*, 30:685–696, 2020.

[40] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021.

[41] Mang Ye, Jianbing Shen, Senior Member, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[42] Rui Yu, Dawei Du, Rodney LaLonde, Daniel Davila, Christopher Funk, Anthony Hoogs, and Brian Clipp. Cascade transformers for end-to-end person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7267–7276, 2022.

[43] Guowen Zhang, Pingping Zhang, Jinqing Qi, and Huchuan Lu. Hat: Hierarchical aggregation transformers for person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 516–525, 2021.

[44] Jialiang Zhang, Lixiang Lin, Jianke Zhu, Yang Li, Yun-chen Chen, Yao Hu, and Steven CH Hoi. Attribute-aware pedestrian detection in a crowd. *IEEE Transactions on Multimedia*, 23:3085–3097, 2020.

[45] Xinyu Zhang, Xinlong Wang, Jia-Wang Bian, Chunhua Shen, and Mingyu You. Diverse knowledge distillation for end-to-end person search. *Proc. AAAI Conference on Artificial Intelligence*, 2021.

[46] Cairong Zhao, Zhicheng Chen, Shuguang Dou, Zefan Qu, Jiawei Yao, Jun Wu, and Duoqian Miao. Context-aware feature learning for noise robust person search. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[47] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[48] Yingji Zhong, Xiaoyu Wang, and Shiliang Zhang. Robust partial matching for person search in the wild. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020.