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.

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

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 mis-aligned 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-
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 \( F \in \mathbb{R}^{h \times w \times c} \), 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 \( F \in \mathbb{R}^{h \times w \times c} \) 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 \( M \in \mathbb{R}^{h \times w \times c} \) 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 $\bar{F} \in \mathcal{R}^{h \times w \times \bar{c}}$ 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 contain 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, metrics, and experimental details. Further, 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.

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 34,334 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 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.
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>CUHK-SYSU</th>
<th>PRW</th>
</tr>
</thead>
<tbody>
<tr>
<td>MGN+OR [39]</td>
<td>53.3</td>
<td>80.9</td>
</tr>
<tr>
<td>QEEPS [29]</td>
<td>49.5</td>
<td>78.7</td>
</tr>
<tr>
<td>AlignPS [38]</td>
<td>45.9</td>
<td>74.9</td>
</tr>
<tr>
<td>PSTR [2]</td>
<td>46.9</td>
<td>76.7</td>
</tr>
<tr>
<td>COAT [42]</td>
<td>47.6</td>
<td>78.7</td>
</tr>
<tr>
<td>Others (SAT)</td>
<td>49.5</td>
<td>78.7</td>
</tr>
<tr>
<td>Others (SAT) + CBGM [26]</td>
<td>50.1</td>
<td>79.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>CUHK-SYSU</th>
<th>PRW</th>
</tr>
</thead>
<tbody>
<tr>
<td>MGN+OR [39]</td>
<td>53.3</td>
<td>80.9</td>
</tr>
<tr>
<td>QEEPS [29]</td>
<td>49.5</td>
<td>78.7</td>
</tr>
<tr>
<td>AlignPS [38]</td>
<td>45.9</td>
<td>74.9</td>
</tr>
<tr>
<td>PSTR [2]</td>
<td>46.9</td>
<td>76.7</td>
</tr>
<tr>
<td>COAT [42]</td>
<td>47.6</td>
<td>78.7</td>
</tr>
<tr>
<td>Others (SAT)</td>
<td>49.5</td>
<td>78.7</td>
</tr>
<tr>
<td>Others (SAT) + CBGM [26]</td>
<td>50.1</td>
<td>79.2</td>
</tr>
</tbody>
</table>

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.
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.

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

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

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.

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


