Knowledge-Based Systems 304 (2024) 112483

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Learning adaptive shift and task decoupling for discriminative one-step person search

Qixian Zhang [a,b], Duoqian Miao [a,b,*], Qi Zhang [a,b], Changwei Wang [c], Yanping Li [a,b], Hongyun Zhang [a,b], Cairong Zhao [a,b]

[a] Department of Computer Science and Technology, Tongji University, Shanghai 201804, China

[b] Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai 201804, China

[c] Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Jinan 250014, China

ARTICLE INFO

Keywords:
Person search
Person re-identification
Discriminative features
Adaptive shift
Task decoupling

ABSTRACT

Mainstream person search models aim to jointly optimize person detection and re-identification (ReID) in a one-step manner. Despite notable progress, existing one-step person search models still face three major challenges in extracting discriminative features: 1) incomplete feature extraction and fusion hinder the effective utilization of multiscale information, 2) the models struggle to capture critical features in complex occlusion scenarios, and 3) the optimization objectives of person detection and ReID are in conflict in the shared feature space. To address these issues, this study proposes a novel adaptive shift and task decoupling (ASTD) method that aims to enhance the accuracy and robustness of extracting discriminative features within the region of interest. In particular, we introduce a scale-aware transformer to handle scale/pose variations and occlusions. This transformer incorporates scale-aware modulation to enhance the utilization of multiscale information and adaptive shift augmentation to learn adaptation to occlusions dynamically. In addition, we design a task decoupling mechanism to hierarchically learn independent task representations using orthogonal loss to decouple two subtasks during training. Experimental results show that ASTD achieves state-of-the-art performance on the CUHK-SYSU and PRW datasets. Our code is accessible at https://github.com/zqx951102/ASTD.

1. Introduction

Person search [1–6] aims to locate and recognize specific query persons in scene images captured using different cameras. It comprises two subtasks, namely, person detection [7,8] and person re-identification (ReID) [9–16]. The detection task detects all persons in a scene image and generates their bounding boxes, and person ReID aims to retrieve specific individuals across multiple cameras for identification. Considering its application to complete-scene images, person search offers greater potential than the ReID task for practical applications, such as criminal tracking, video surveillance, and smart cities.

Current works can be categorized into two-step and one-step methods according to the training approach. Two-step methods [3,17–20] typically use two independent models to process the two subtasks sequentially. Initially, an off-the-shelf detector is used to locate persons, and then, cropped person patches are fed into a ReID model for identification. However, the use of two separate networks to process the two subtasks is time-consuming and resource-intensive. By contrast, one-step methods [2,21,22] achieve higher efficiency by simultaneously learning person detection and recognition in a shared underlying framework. Given an uncropped image, the model identifies and predicts bounding boxes and identity features for all detected individuals in a single network pass.

Despite significant progress, existing one-step methods still face three major challenges in extracting discriminative features to distinguish individuals accurately:

(**CH1**) Significant scale and pose variations complicate person identification. As presented in Fig. 1(a), pedestrian size distribution varies greatly in real-world scenes because of diverse camera viewpoints and positions. The statistical analysis of the PRW dataset [3] reveals that 23.5% of pedestrians exhibit scale variation issues. To address this, some methods [23,24] use fixed-scale convolutional layers to generate multiscale features. However, these methods are limited in their flexibility to effectively capture and fuse multiscale features. Some models use feature pyramids [25,26], but inconsistent features and simple

[*] Corresponding author at: Department of Computer Science and Technology, Tongji University, Shanghai 201804, China.
E-mail addresses: zhangqx@tongji.edu.cn (Q. Zhang), dqmiao@tongji.edu.cn (D. Miao), zhangqi_cs@tongji.edu.cn (Q. Zhang), wangchangwei2019@ia.ac.cn (C. Wang), liyp8023yz@tongji.edu.cn (Y. Li), zhanghongyun@tongji.edu.cn (H. Zhang), zhaocairong@tongji.edu.cn (C. Zhao).

https://doi.org/10.1016/j.knosys.2024.112483
Received 2 July 2024; Received in revised form 22 August 2024; Accepted 4 September 2024
Available online 11 September 2024

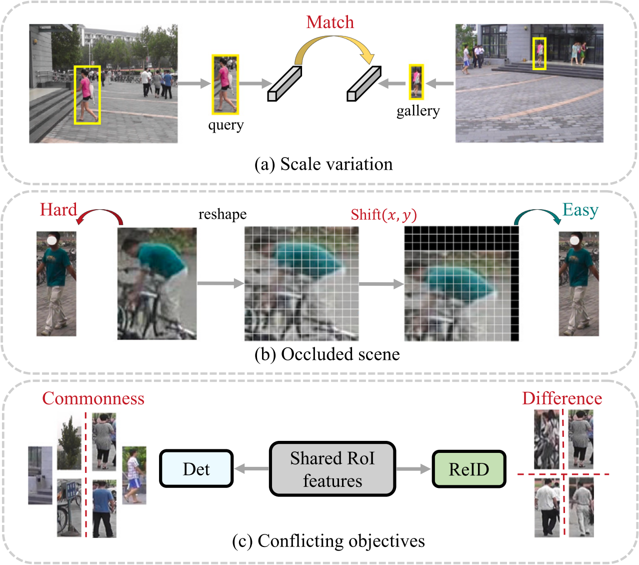


Fig. 1. Illustrations of major challenges in one-step person search: (a) Scale variation: pedestrians exhibit varying appearances and scales in images due to the diverse viewpoints and flexible positions of cameras. (b) Occluded scene: occlusion by other people or background objects leads to disordered and incomplete features within the RoI, reducing their quality. (c) Conflicting objectives: the detection task attempts to find the commonality of all individuals, and the ReID task emphasizes the difference between multiple instances.

fusion strategies can introduce background noise, which degrades ReID performance.

(**CH2**) Occlusion by other people or objects degrades the quality of the RoI features and complicates the matching process, as shown in Fig. 1 (b). To address this, some graph-based methods [27,28] model the graph topology of 14 pedestrian nodes, which increases computational complexity. He et al. [29] shuffled and regrouped pedestrian part embeddings to produce robust ReID features, and Yu et al. [23] exchanged tokens of partial instances to simulate real occlusion scenarios, thereby enhancing robustness. Although these methods show advancements in handling partial occlusion, they show limited unpromising adaptability for complex occlusion scenarios.

(**CH3**) Optimization conflicts between person detection and person ReID cause difficulties in learning optimal task-shared features. As shown in Fig. 1(c), pedestrian detection focuses on learning commonalities to distinguish pedestrians from the background, whereas ReID attempts to differentiate multiple identities. Han et al. [30,31] simplified the one-step pipeline by treating RoI features as specific to ReID and eliminating detection loss in the ReID head, thereby resulting in fewer and less precise training bounding boxes. Chen et al. [22] proposed norm-aware embedding (NAE), disintegrating embeddings into a norm and an angle for detection and ReID, respectively. This method has been adopted in various studies [23,32], but excessive parameter sharing leads to suboptimal performance in handling the two distinct tasks.

Considering the above discussion, we propose a novel framework called ASTD for learning discriminative features in a one-step person search. This framework incorporates the merits of CNNs and transformers to address the above key issues in an end-to-end manner. Inspired by Cascade R-CNN [33], we adopt a three-stage structure to refine the detection and ReID features from coarse to fine. In particular, we introduce a scale-aware transformer (SAT) at every stage for highly discriminative feature embedding. SAT processes backbone features using a convolutional encoder to capture local features. Its intermediate representation is divided into two parts: one part aggregates scale information within the RoI using SAM. It equips the framework with multiscale adaptability by dynamically adjusting the convolutional kernel size and fusing channel-level information to handle scale/pose variations (**CH1**). The other part is processed by adaptive shift augmentation (ASA), which transforms shift operations into a differentiable classification problem using Gumbel Softmax sampling for learning dynamic shifts. This design improves the framework's effectiveness in handling occlusion issues (**CH2**). In addition, we introduce a task decoupling mechanism (TDM) with an orthogonal loss. TDM initially learns shared invariant features and then uses attention mechanisms for hierarchical fine-grained decoupling. The orthogonal loss ensures feature discriminability and subtask independence during training, avoiding conflicting objectives between subtasks (**CH3**). Experimental results show that ASTD attains state-of-the-art outcomes on the CUHK-SYSU and PRW datasets.

Our primary contributions are outlined below:

(1) We propose a novel one-step person search framework called ASTD that is devoted to learning highly discriminative features to effectively improve accuracy and robustness.

(2) We propose a SAT that allows multiscale feature extraction and ASA to handle scale/pose variations and occlusions, respectively.

(3) We present a novel TDM with an elaborated orthogonal loss to decouple person detection and person ReID, effectively improving task-shared features.

(4) Extensive experiments on two benchmark datasets demonstrate the superiority of our method, which outperforms existing state-of-the-art methods. In particular, on the PRW dataset, it improves the top-1 score by 3.2% over the baseline.

The remainder of this article is structured as follows: Section 2 reviews the related work, Section 3 explains our methodology in detail, Section 4 presents the experiments and analyzes the results, and Section 5 concludes the article.

2. Related works

2.1. Person search

Existing person search methods are typically classified into two categories based on their training steps.

(1) Two-step methods: These methods divide person search into two subtasks, namely, person detection and person ReID, and train them using two independent models. Zheng et al. [3] first assessed different pairings of detectors and ReID networks. Lan et al. [34] proposed a cross-level semantic alignment (CLSA) approach to address the multiscale matching problem. Chen et al. [17] revealed the conflicting target problem and obtained more distinctive features using a dual-stream model. Wang et al. [19] further observed the inconsistency in training ReID models on hand-drawn images and mitigated this issue by generating query-like bounding boxes and using detected bounding boxes for training. Yao et al. [20] introduced an OR similarity that considers the objectness and repulsion information.

(2) One-step methods: These methods develop a unified model for the end-to-end training of the two subtasks. In general, this approach has fewer parameters and higher efficiency. Xiao et al. [2] initially proposed an end-to-end approach based on Faster R-CNN, sharing the base layers with ReID networks. Chang et al. [35] proposed a relational context-aware agent (RCAA) approach to efficiently localize target persons in whole scene images without bounding box annotations. Dong et al. [36] designed a bidirectional interaction network to alleviate redundant contextual information outside the bounding box. Several methods were considered utilizing query images. Munjal et al. [37] were the first to implement a query-guided mechanism to iteratively narrow down the search area. Yan et al. [25] presented an anchor-free person search network and addressed the alignment problems at three levels (scale, region, and task). Li and Miao [32] proposed a SeqNet, which sequentially used two Faster R-CNN networks for detection and ReID. Jaffe et al. [38] reduced the size of the search

gallery by reducing similar scenes, saving computational resources. Song et al. [39] proposed to improve the quality of person bounding boxes by considering the interactions between persons and scenes. Recently, some approaches [40,41] have combined large language models (LLMs) and diffusion models for text and image-based person search. This integrated strategy opens up new research directions in the field of person search.

As highlighted by [17], pedestrian detection emphasizes learning the common characteristics of all individuals, whereas ReID attempts to identify differences among multiple identities. Han et al. [30,31] addressed this issue using point-based spatial sampling to obtain RoI features tailored specifically for the ReID task. Chen et al. [22] disintegrated the feature vectors into a norm and an angle to separately assess detection confidence and identity similarity. Differing from these methods, we identify that the differing demands for features are the core reason for the conflict and hinder discriminative feature learning.

2.2. Feature learning in occlusion

Recent pose-guided methods have been developed to address issues in person ReID. Wang et al. [28] used a graph-based approach to capture topological information and learn the correspondence between nodes. Wang et al. [27] used pose information to distinguish between unoccluded and occluded features. However, this significantly increases computational complexity. Dou et al. [42] integrated input semantic information from the ReID network into a guided alignment module to obtain features of the foreground and body parts. He et al. [29] proposed rearranging patch embeddings through shift and patch shuffle operations to produce robust ReID features. Li et al. [43] introduced a part discovery technique that uses part-aware transformers to address occlusion issues in person ReID. Li et al. [10] dynamically adjusted weights based on information entropy to reduce the uncertainty caused by occlusions.

Research on occlusion in person search is limited. Yu et al. [23] proposed randomly mixing partial tokens to simulate occlusion attention, but this might lead to learning incorrect features if the tokens do not reflect the actual scene. Fiaz et al. [24] enhanced single-shift operations after feature grouping, yet these may lack flexibility for extensive occlusions. Zhang et al. [44] introduced an attentive multi-granularity perception module, addressing appearance variations and occlusions within the RoI. In this study, we present a novel ASA strategy that dynamically learns and adjusts end-to-end feature map positions to mitigate scene occlusions effectively.

2.3. Transformer-based approaches

Since the application of the VIT model [45] in image recognition tasks, it has been widely used in various computer vision fields, including person ReID and person search. He et al. [29] proposed reordering patch embeddings through shift and patch shuffle techniques in a pure transformer, generating more robust ReID features. Wang et al. [46] introduced the neighbor transformer, leveraging neighbor features to obtain robust feature representations for person ReID. Li et al. [47–49] proposed a multi-granularity cross transformer, that gradually learns the significant features of various local structures of a person in a global context. Recently, PSTR [50] and cascade occluded attention transformer (COAT) [23] were introduced in the field of person search, both integrating transformers and achieving good performance. PSTR, which is built on the DETR [51] framework, designs a detection encoder–decoder for pedestrian detection and a distinctive ReID decoder. On the basis of Cascade R-CNN, COAT uses multiscale convolutional transforms in multiple stages to force the network to learn scale/occlusion feature representations from coarse to fine. In recent years, transformer-based approaches have shown strong capabilities in text-based and image-based person search tasks. TBPS-CLIP [40] uses cross-modal contrastive learning using a transformer architecture for an efficient text-person image matching. DP [41] combines transformer-based LLMs to generate high-quality text annotations consistent with image contents.

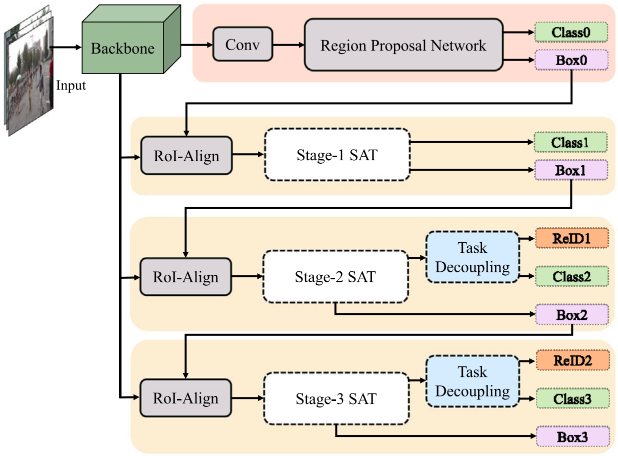


Fig. 2. The architecture of the ASTD framework starts by processing the input image using a backbone network to extract stem features. These features are then fed through a convolutional layer to the region proposal network (RPN). The backbone features and bounding boxes are subsequently forwarded to various stages to acquire RoI-Align pooled features. Each stage uses the bounding boxes estimated from the previous stage, except for Stage-1, which uses the bounding boxes generated by the PRN. The framework includes several key components: (1) the scale-aware transformer (SAT) module, designed to address scale/pose variations and occlusions, and (2) the task decoupling mechanism (TDM) in Stage-2 and Stage-3 to address conflicts between the detection and ReID tasks.

By contrast, we design a context aggregator that leverages the strengths of CNNs and transformers to address scale/pose variations at each stage. In addition, compared with a fixed-scale feature extractor, we dynamically adjust the size of the heads to extract scale information and fuse it at the channel level, thereby enhancing both utilization and adaptability to multiscale information.

3. Methodology

3.1. Overall architecture

The proposed ASTD builds upon the COAT [23], a state-of-the-art framework for person search that excels in handling challenging scenarios, such as occlusions and scale variations. COAT is particularly notable for its occluded attention transformer mechanism, which simulates real-world occlusions by strategically masking or exchanging partial tokens among instances within a mini-batch. This approach compels the model to learn more robust and discriminative embeddings, enhancing its ability to accurately identify persons even under occlusion. In addition, COAT's three-stage cascade design supports a coarse-to-fine learning process, progressively refining both detection accuracy and ReID performance. This hierarchical refinement ensures that the final output is highly precise, making COAT an ideal baseline for our enhancements.

In ASTD, we further enhance COAT's architecture by introducing a context aggregator at each stage that leverages the strengths of both CNNs and transformers, as shown in Fig. 2. To address scale/pose variations and occlusions, we incorporate a SAT module that progressively learns discriminative features. In addition, a TDM is implemented to address the conflicting objectives between detection and ReID.

The framework starts by extracting 1024-d features from a ResNet50 backbone, followed by generating region proposals with the RPN. Each proposal is processed using RoI-Align [52] to pool base feature maps. A multistage cascade strategy then refines representations for more precise detection and ReID. Optimization occurs in the first stage with class and box heads and in the subsequent two stages with extra ReID heads, building upon the prior stage's bounding box estimates.

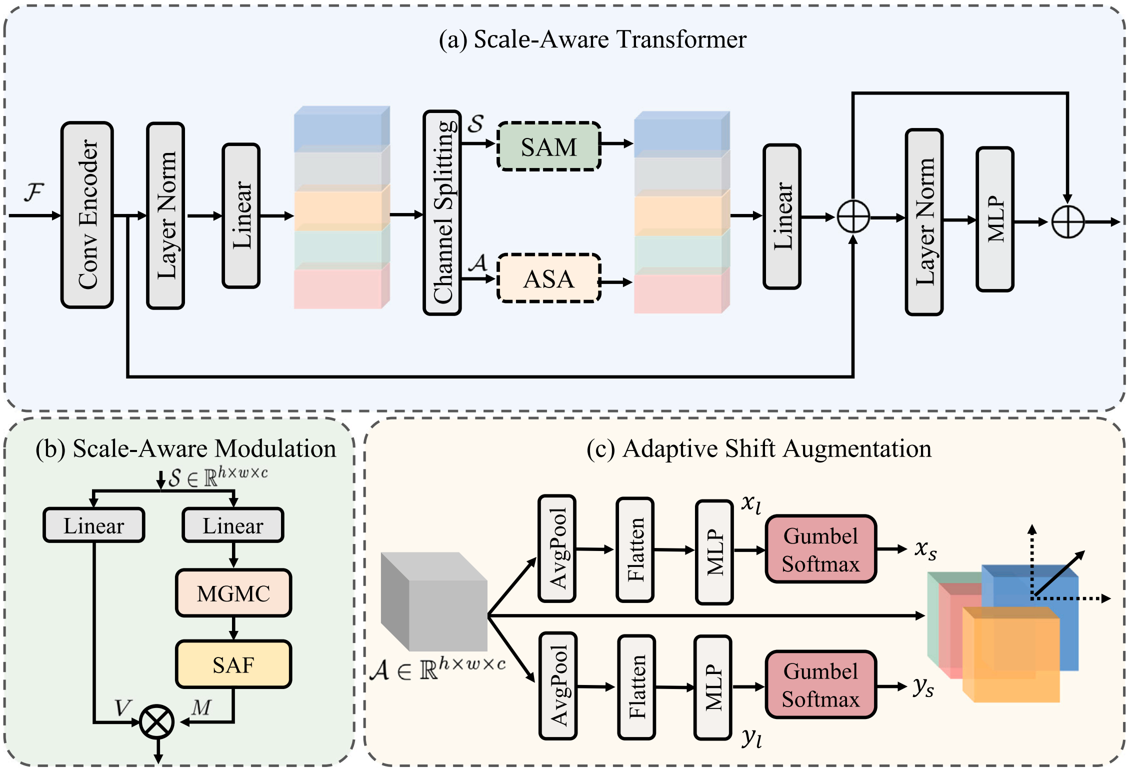


Fig. 3. (a) The detailed architecture of the presented scale-aware transformer (SAT). The input feature $\mathcal{F}$ initially goes through a convolutional encoder to capture fine-grained features. After layer normalization and linear transformation, the features double in size and split into two parts: $\mathcal{S} \in \mathbb{R}^{h\times w\times c}$ goes to (b) scale-aware modulation (SAM) for learning multiscale information via MGMC and SAF, and $\mathcal{A} \in \mathbb{R}^{h\times w\times c}$ goes to (c) adaptive shift augmentation (ASA) for addressing occlusions by generating offsets x_s and y_s using MLP and the Gumbel Softmax function. Finally, the outputs from SAM and ASA are fused, undergo further linear transformation, and are added to the initial convolutional output before passing through a normalization layer and an MLP block.

3.2. Scale-aware transformer

Existing methods struggle to adapt to the scale/pose variations and occlusions that pedestrians often experience. They typically use fixed-scale feature extractors that fail to effectively capture varied pedestrian features or dynamically adjust to specific occlusions. To enhance adaptability, we introduce the SAT, a context aggregator that leverages CNNs and transformers for dynamic adaptation to scale and occlusion variations. SAT processes stem features using a convolutional encoder to extract local features and then uses a transformer for global relational inference.

As depicted in Fig. 3(a), the feature tensor $\mathcal{F} \in \mathbb{R}^{h\times w\times c}$ enters the Conv Encoder, followed by normalization and a linear layer that doubles the channels and splits the output into two halves. One part, $\mathcal{S} \in \mathbb{R}^{h\times w\times c}$, is processed by SAM to learn scale and pose variations, and the other part, $\mathcal{A} \in \mathbb{R}^{h\times w\times c}$, is used by ASA to enhance robustness against occlusions. The outputs of SAM and ASA are combined and then fused with the Conv encoder [53] output via a skip connection, further mixed channel-wise using a normalization layer and an MLP block. The final SAT output undergoes global average pooling before being directed to various heads.

3.3. Scale-aware modulation

The SAM module is proposed to address scale/pose variations of pedestrians within the RoI, as shown in Fig. 3(b). This module increases the model's flexibility in handling multiscale information through multi-granularity mixed convolution (MGMC) and scale-aware fusion (SAF).

3.3.1. MGMC

As illustrated in Fig. 4(a), convolutional layers with diverse kernel sizes are used to capture spatial information across various scales within the RoI. Increasing the N head parameter extends the receptive field, enhancing the capability of the model to capture discriminative features. The implementation process of MGMC proceeds as follows:

$$MGMC(\mathcal{S}) = \left[DW_{k_1\times k_1}(s_1), \dots, DW_{k_N\times k_N}(s_N) \right] \tag{1}$$

where $s \in \{s_1, s_2, \dots, s_N\}$ indicates that the input feature $\mathcal{S}$ is divided into N heads along the channel dimension, $k_i \in \{3, 5, \dots, K\}$ indicates that the convolution kernel size increases by 2 per head, DW denotes depthwise convolution, and [] signifies concatenation along the channel dimension.

3.3.2. SAF

A lightweight and efficient SAF module is developed to improve the adaptability and robustness of the model across multiple scales, as shown in Fig. 4(b). The implementation of the SAF proceeds as follows:

$$\begin{aligned} H_j^i &= DW_{k_j\times k_j}\left(s_j^i\right) \in \mathbb{R}^{h\times w\times 1} \\ G_i &= W_{\text{intra}}\ \left(\left[H_1^i, H_2^i, \dots, H_N^i\right]\right) \\ M &= W_{\text{inter}}\ \left(\left[G_1, G_2, \dots, G_G\right]\right) \end{aligned} \tag{2}$$

where W_{intra} and W_{inter} represent the weights of the pointwise convolution. $i \in \{1, 2, \dots, G\}$ and $j \in \{1, 2, \dots, N\}$, where $G = \frac{c}{N}$ and N, respectively, denote the number of groups and heads. Here, c denotes the number of channels. $H_j \in \mathbb{R}^{h\times w\times G}$ represents the j-th head with depthwise convolution, and H_j^i represents the i-th channel in the j-th head.

After extracting multiscale information with MGMC and fusing it with SAF, an output feature map M is generated to modulate V through

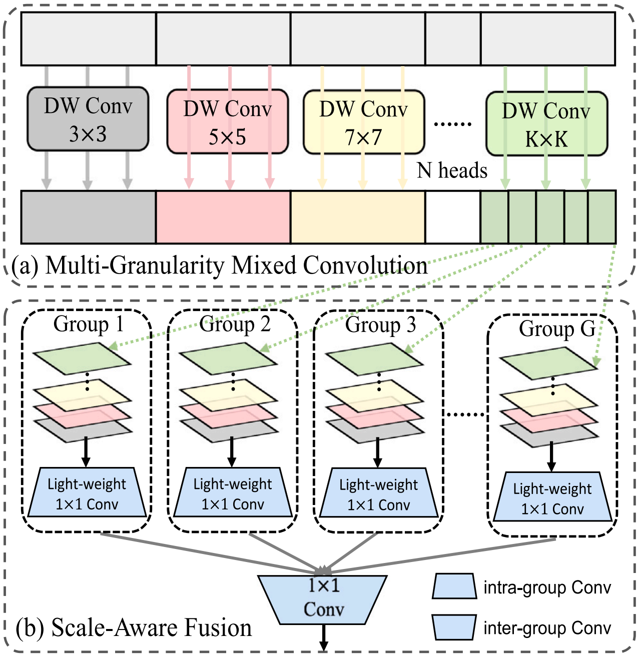


Fig. 4. Implementation process of MGMC and SAF. (a) MGMC divides input channels into N heads for independent depthwise convolutions, capturing multiscale information. (b) SAF reorganizes features by selecting channels to form new groups, enhancing diversity. It then applies pointwise convolution for lightweight intragroup and intergroup fusion.

a dot product operation. The output feature O is computed from the input feature $\mathcal{S}$ as follows:

$$\begin{aligned} V &= W_v \mathcal{S} \\ M &= SAF\left(MGMC\left(W_m \mathcal{S}\right)\right) \\ O &= M \otimes V \end{aligned} \tag{3}$$

where W_v and W_m represent the weights of the linear layer and $\otimes$ represents the dot product operation.

3.4. Adaptive shift augmentation

To enhance robustness against occlusions within the RoI, Yu et al. [23] exchanged parts of pedestrians within a mini-batch, which might lead to inaccurate part information across instances. By contrast, our ASA module shown in Fig. 3(c) innovatively addresses occlusions by transforming traditional, nondifferentiable shift operations into a differentiable classification problem. This approach not only enables trainable shift operations but also significantly enhances the capacity to recognize occluded features.

The ASA module processes the tensor $\mathcal{A} \in \mathbb{R}^{h \times w \times c}$ via adaptive average pooling and flattening along the x-axis and y-axis, respectively. Subsequently, it uses separate MLPs to directly produce logits. The formulation is as follows:

$$\begin{aligned} x_l &= MLP_x(Flatten(AvgPool(\mathcal{A},(1,h)))) \\ y_l &= MLP_y(Flatten(AvgPool(\mathcal{A},(w,1)))) \end{aligned} \tag{4}$$

Gumbel Softmax sampling [54] is performed on these predicted logits. Soft sampling is opted here, producing a probability distribution rather than discrete samples. This approach facilitates the calculation of shift values for x and y by identifying positions with the maximum probability.

$$\begin{aligned} x_s &= Argmax\left(GumbelSoftmax\left(x_l\right)\right) - \delta \\ y_s &= Argmax\left(GumbelSoftmax\left(y_l\right)\right) - \delta \end{aligned} \tag{5}$$

where x_s and y_s represent the shift values in the x-axis and y-axis directions, respectively, with $s \in \{1, 2, \ldots, g\}$, where g denotes the number of groups. $Argmax$ denotes the operation used to determine the location of the maximum value. The parameter δ regulates the probability mapping range to prevent exceeding the target area.

Partition the feature $\mathcal{A}$ into g groups according to the channel dimension, denoted as $\{\mathcal{A}_\gamma\}_{\gamma=1}^{g}$ for each group, where $\mathcal{A}_\gamma \in \mathbb{R}^{h \times w \times c/g}$. The formula for the shift operation is as follows:

$$\begin{aligned} \mathcal{A}_\gamma[0:h,:,:] &\leftarrow \mathcal{A}_\gamma[0+y_s:h+y_s,:,:] \\ \mathcal{A}_\gamma[:,0:w,:] &\leftarrow \mathcal{A}_\gamma[:,0+x_s:w+x_s,:] \end{aligned} \tag{6}$$

When $x_s, y_s > 0$, channel features are shifted to the upper right, similarly for other cases. After shifting, out-of-range pixels are discarded, and the resulting empty spaces are filled with zeros. Notably, the shift operation is efficient and lightweight, requiring only memory copying without additional parameters.

3.5. Task decoupling mechanism

To mitigate the conflicting objectives between subtasks, our strategy emphasizes independence and orthogonality in the task optimization process. This approach is in contrast with NAE-based methods [22,23], which decompose feature embeddings into a norm and an angle within a polar coordinate system. By enhancing independence and orthogonality, we effectively decouple the subtasks, ensuring that the optimization of each task does not negatively affect the others. As shown in Fig. 5, the task-invariant features F_{task}^{k} are computed as follows:

$$F_{task}^{k} = ReLU\left(Conv_k\left(F_{task}^{k-1}\right)\right) \tag{7}$$

$$w_s = Softmax(\omega), \quad \omega \sim \mathcal{N}(0,1) \tag{8}$$

$$F_s = w_s \times F_{task}^{k} \tag{9}$$

where $k \in \{1, 2, \ldots, M\}$ and $Conv_k()$ denotes the k-th convolutional layer. F_s represents the task-specific features for $s \in \{1, 2, \ldots, n\}$ (with $n = 2$ for classification and detection), and w_s is the attention weight distribution derived from the initialized weight matrix ω.

Orthogonality loss $\mathcal{L}_{ort}$ is applied to supervise the initial weights, minimizing their correlation and fostering more independent representations for each task:

$$\mathcal{L}_{ort} = \left\| w_1^T \times w_2 \right\|_F = \sum_{i=1}^{\psi} \sum_{j=1}^{\varphi} \left(w_1^T w_2\right)_{ij}^2 \tag{10}$$

where $\|\cdot\|_F$ represents the Frobenius norm and ψ and φ denote the dimensions of the corresponding weight matrices, respectively. Here, w_1 and w_2 represent the weight vectors specific to different tasks. The orthogonality loss aims to minimize the norm of the dot product between task-specific weight vectors, thus encouraging their divergence in the feature space. This encourages task-specific weight vectors to be orthogonal to each other, thereby enhancing their independence in capturing distinct features.

Through backpropagation, weights are incrementally adjusted to enhance orthogonality, ensuring that attention weights for each task remain distinct throughout the training process.

3.6. Loss function

During training, the ASTD network is supervised using the detector loss $\mathcal{L}_{det}$, the online instance matching (OIM) loss $\mathcal{L}_{oim}$, and the orthogonality loss $\mathcal{L}_{ort}$.

Detector loss $\mathcal{L}_{det}$: The loss includes box regression and classification losses from the PRN and the three SAT stages.

$$\mathcal{L}_{det} = \mathcal{L}_{cls} + \mathcal{L}_{reg} \tag{11}$$

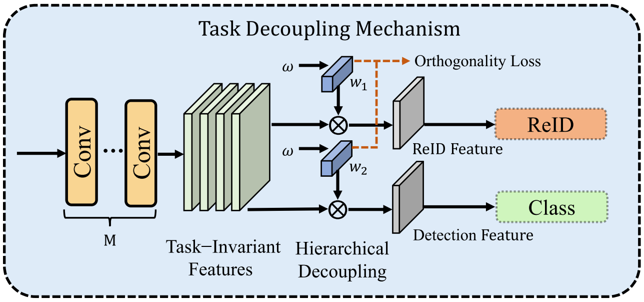


Fig. 5. An illustration of the proposed task decoupling mechanism. Task-invariant features are extracted using a series of 1 × 1 convolutional layers. Subsequently, these features are decoupled into task-specific features using two separate attention mechanisms. In addition, an orthogonality loss supervises the attention weights to maintain task independence during training.

where $\mathcal{L}_{cls}$ calculates the cross-entropy loss for the predicted classification probabilities of the estimated boxes and $\mathcal{L}_{reg}$ represents the Smooth-L1 loss between the regression vectors of the ground truth and the foreground boxes.

OIM loss: To supervise person ReID, similar to other person search methods [23], the OIM loss [2] is used.

$$p_z = \frac{\exp\left(v_i^Z x/\tau\right)}{\sum_{j=1}^{L}\exp\left(v_j^Z x/\tau\right) + \sum_{k=1}^{Q}\exp\left(u_k^Z x/\tau\right)} \tag{12}$$

$$\mathcal{L}_{oim} = E_x\left[\log p_z\right] \tag{13}$$

where $z \in \{1, 2, \dots, L\}$, z indicates the label of input feature x, and p_z indicates the probability that x belongs to label z. τ is a hyperparameter. v^Z is the primary feature vector of labeled persons stored in a circular queue of size L. u^Z is the feature vector of unlabeled IDs stored in a lookup table of size Q to store unlabeled feature vectors.

The proposed ASTD is trained using a multistage loss function.

$$\mathcal{L}_{all} = \sum_{t=0}^{T} \mathcal{L}_{det}^{t} + \mathbb{I}(t > 1)\left(\mathcal{L}_{oim}^{t} + \mathcal{L}_{ort}^{t}\right) \tag{14}$$

where $t \in \{0, 1, \dots, T\}$ signifies the index of the stage, T represents the count of cascaded stages, and the indicator function $\mathbb{I}(t > 1)$ indicates that ReID loss and orthogonality loss are not considered in the first stage.

For clarity, we present the detailed training process of the ASTD framework in Algorithm 1.

4. Experiments

In this section, we start by providing a detailed description of the two datasets and their settings. Next, we compare the performance of our approach with state-of-the-art methods. Then, we perform ablation experiments to assess the contribution of each module. Finally, we provide more performance and qualitative results to further demonstrate the effectiveness of our approach.

4.1. Datasets and settings

4.1.1. Datasets

In our experiments, we utilize two benchmark datasets: CUHK-SYSU and PRW. CUHK-SYSU [2] is a large-scale dataset consisting of 18,184 scene images primarily sourced from real-world street cameras and movie screenshots, with a total of 96,143 annotated bounding boxes and 8432 identities. The dataset is divided into two non-overlapping parts. The training set includes 11,206 images, 55,272 annotated bounding boxes, and 5532 identities. The test set comprises 6,978 images, covering 2900 pedestrians, with 40,871 labeled pedestrians.

Algorithm 1 Training Process of ASTD

```
Input: Training set $I$, total epochs $epochs$, iterations $T$
Output: Trained Model weight $\mathbb{W}$
Initialize the model weight $\mathbb{W}$
for $e = 1$ to $epochs$ do
    for each batch $B$ sampled from $I$ do
        Extract pedestrian features $F$ through Backbone and RoI-Align from $B$
        Calculate RPN loss $\mathcal{L}_{det}$ by Eq (11)
        for $t = 1$ to $T$ do
            if $t = 1$ then
                Feed $F$ into the Scale-Aware Transformer and optimize using class and box heads
                Compute the detector loss $\mathcal{L}_{det}$ by Eq (11)
            else
                Use bounding box regression estimates from the previous stage
                Refine $F$ through SAT by Eqs. (1) – (6)
                Decouple task features by Eqs. (7) – (9)
                Calculate orthogonal loss $\mathcal{L}_{ort}$ by Eq (10)
                Compute losses: detector loss $\mathcal{L}_{det}$ by Eq (11), OIM loss $\mathcal{L}_{oim}$ by Eqs. (12) and (13)
            end if
            Calculate total loss $\mathcal{L}_{all}$ by Eq (14)
            Back propagate to update $\mathbb{W}$
        end for
    end for
end for
return trained model weight $\mathbb{W}$
```

PRW [3] is extracted from six static cameras on a university campus, exhibiting various camera styles and significant scale variations. It comprises a total of 11,816 video frames, with 34,304 bounding boxes annotated for 932 labeled identities. The training set contains 5704 images, 18,048 pedestrians, and 482 identities. The test set includes 6112 images and 2057 query persons, covering 450 identities.

4.1.2. Evaluation protocols

We use the same evaluation metrics as those used in earlier works [2], incorporating mean average precision (mAP) and top-1 score as evaluation metrics. A box is considered a match if the overlap ratio between the predicted box and the ground truth box of the same identity is greater than 0.5 IoU.

4.1.3. Implementation details

Our model is built using the PyTorch framework and operates on an NVIDIA Tesla V100 GPU. During training, the IoU thresholds for detection across the three stages are set to 0.5, 0.6, and 0.7. An SGD optimizer with a momentum of 0.9 and weight decay of 5×10^{-4} is used for training over 15 epochs with a batch size of 3 and a learning rate of 0.003. We perform warm-up at the first epoch and decrease at the 11th epoch. During testing, nonmaximum suppression (NMS) is used with thresholds set to 0.4, 0.4, and 0.5 to eliminate redundant bounding boxes.

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

In this section, we evaluate our approach against the state-of-the-art techniques using 2 benchmark datasets, including 7 two-step methods and 15 one-step methods.

Table 1
Comparison of mAP and top-1 score with state-of-the-art methods on two benchmark datasets. The highest score is emphasized in bold.

Method	Ref	Backbone	CUHK-SYSU		PRW	
			mAP	top-1	mAP	top-1
Two-step methods						
IDE [3]	CVPR17	ResNet50	–	-	20.5	48.3
MGTS [17]	ECCV18	VGG16	83.0	83.7	32.6	72.1
CLSA [34]	ECCV18	ResNet50	87.2	88.5	38.7	65.0
RDLR [18]	ICCV19	ResNet50	93.0	94.2	42.9	70.2
IGPN [55]	CVPR20	ResNet50	90.3	91.4	47.2	87.0
TCTS [19]	CVPR20	ResNet50	93.9	95.1	46.8	87.5
OR [20]	TIP21	ResNet50	92.3	93.8	52.3	71.5
One-step with CNNs						
OIM [2]	CVPR17	ResNet50	75.5	78.7	21.3	49.4
RCAA [35]	ECCV18	ResNet50	79.3	81.3	–	–
CTXG [56]	CVPR19	ResNet50	84.1	86.5	33.4	73.6
NAE [22]	CVPR20	ResNet50	91.5	92.4	43.3	80.9
AlignPS+ [25]	CVPR21	ResNet50-DCN	94.0	94.5	46.1	82.1
SeqNet [32]	AAAI21	ResNet50	94.8	95.7	47.6	87.6
CANR+ [57]	TCSVT22	ResNet50	93.9	94.5	44.8	83.9
SPG [39]	TII24	ResNet50	95.0	95.9	48.4	89.8
SeqNeXt+GFN [38]	WACV23	ResNet50	94.7	95.3	51.3	**90.6**
DMRNet++ [30]	TPAMI23	ResNet50	94.5	95.7	52.1	87.0
One-step with Transformers						
PSTR [50]	CVPR22	ResNet50	93.5	95.0	49.5	87.8
PSTR [50]	CVPR22	PVTv2-B2	95.2	96.2	56.5	89.7
SAT [24]	WACV23	ResNet50	95.3	96.0	55.0	89.2
SOLIDER [58]	CVPR23	Swin -S	95.5	95.8	**59.8**	86.7
COAT [23]	CVPR22	ResNet50	94.2	94.7	53.3	87.4
ASTD (Ours)	-	ResNet50	95.8	96.2	55.7	90.2
ASTD (Ours)	-	ResNet50-DCN	96.0	96.0	56.8	88.5
ASTD (Ours)	-	PVTv2-B2	96.2	96.4	58.5	90.1
ASTD (Ours)	-	Swin -S	**96.4**	**96.7**	59.2	**90.6**

Table 2
Comparison of complexity and inference time for person search approaches on the PRW dataset.

Methods	Params (M)	FLOPs (G)	Time (ms)	mAP
NAE+ [22]	**33**	575	98	44.0
SeqNet [32]	48	550	86	46.7
AlignPS [25]	42	380	**61**	45.9
COAT [23]	37	473	90	53.3
ASTD (Ours)	43	**348**	96	**55.7**

4.2.1. Comparison on the CUHK-SYSU dataset

The results are presented in Table 1. ASTD achieves the best mAP of 95.8% and a top-1 score of 96.2%, outperforming most one-step methods and significantly outperforming two-step methods. The best two-step approach TCTS [19] achieves an mAP score of 93.9%. Using the ResNet50 backbone, ASTD achieves an improved the mAP and top-1 score by 1.6% and 1.5%, respectively, compared with the baseline method COAT [23]. Compared with SOLIDER [58], which considers semantic controllable self-supervised learning, ASTD demonstrates slightly better performance. In addition, using the Swin-S backbone, ASTD establishes a new standard in performance. These improvements are attributed to our successful learning of discriminative features that handle complex occlusions and scale variations. However, in the CUHK-SYSU dataset, the limited number of images per scene and fewer instances of scale variations and occlusions mean that the discriminative features learned are less robust than those in the PRW dataset.

We also evaluate our ASTD against the latest methods on the CUHK-SYSU test set using gallery sizes that range from 50 to 4000. As illustrated in Fig. 6, the performance curves of all compared methods decrease as the gallery size increases. This decrease is primarily due to the challenge of accounting more distracting persons. Nonetheless, our ASTD consistently outperforms both two-step and one-step methods, validating the scalability and robustness of our model in larger search scenarios.

4.2.2. Comparison on the PRW dataset

Table 1 shows ASTD's performance on the PRW dataset, which presents more challenges than the CUHK-SYSU dataset because of its larger gallery size and the presence of many similarly appearing identities. On the ResNet50 backbone, ASTD outperforms the best two-step method OR [20] by 3.4% in mAP, achieving 55.7% mAP and 90.2% top-1 score. It also excels among one-step methods, such as AlignPS and SeqNet, and shows a 6.2% mAP improvement over PSTR [50], which uses a more robust DETR detector. Regarding top-1 score, ASTD matches leading methods SeqNeXT+GFN [38] and SPG [39]. With the Swin-S backbone, ASTD achieves 59.2% mAP and 90.6% top-1 score, significantly outperforming all competitors. This success is attributed to our SAT, which effectively uses discriminative multiscale information.

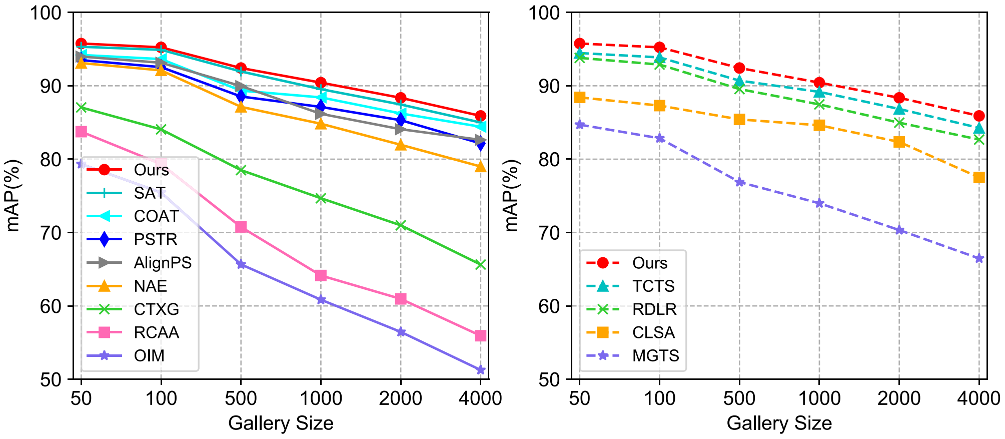
4.2.3. Efficiency comparison

To illustrate the efficiency of our ASTD, we present the inference time (ms) calculated on a Tesla V100 GPU. For a fair comparison, the size of the input image is resized to 900 × 1500 for testing. As depicted in Table 2, our ASTD achieves the lowest FLOPs, primarily due to its lightweight modules and task decoupling strategy. Although our method is slightly slower than COAT, it significantly improves the mAP by 2.4%, demonstrating its practical applicability. Overall, the significant performance improvement of our ASTD, along with its lowest FLOPs, far outweighs the minor increase in inference time.

4.3. Ablation study

In this section, we conduct a detailed ablation study to examine our design choices. First, we evaluate the effectiveness of each component in ASTD and assess the contributions of different stages, validating the decoupling approach. In addition, we evaluate the effectiveness of context aggregators and compare our method with other augmentation mechanisms.

For a fair comparison and to save computational resources and time, we choose ResNet50 as the backbone for subsequent experiments.



(a) One-step methods

(b) Two-step methods

Fig. 6. Comparison with other methods under different gallery sizes on the CUHK-SYSU dataset. (a) shows the comparison with one-step methods, and (b) displays the comparison with two-step methods.

Table 3
Ablation study on gradually adding our contributions to the baseline on the PRW dataset. “GS” denotes Gumbel Softmax.

Baseline	SAM		ASA		TDM		ReID		Detection	
	MGMC	SAF	w/o GS	w/GS	w/o $\mathcal{L}_{ort}$	w/$\mathcal{L}_{ort}$	mAP	top-1	Recall	AP
✓							51.3	87.4	94.6	92.5
✓	✓						52.7	88.1	95.1	92.7
✓		✓					52.1	87.8	94.8	92.6
✓	✓	✓					53.5	88.5	95.3	93.1
✓	✓	✓	✓				53.9	88.7	95.5	93.1
✓	✓	✓		✓			54.8	89.3	**95.8**	92.7
✓	✓	✓		✓	✓		55.2	89.6	95.2	93.0
✓	✓	✓		✓		✓	**55.7**	**90.2**	95.5	**93.4**

Table 4
Comparison with different stage variants of ASTD, where “✓” indicates the use of the proposed SAT and “†” indicates the use of the proposed TDM.

Stage-1	Stage-2	Stage-3	mAP	top-1
(a) with Scale-Aware Transformer:				
✓			52.5	87.6
✓	✓		53.9	88.6
✓	✓	✓	54.8	89.3
(b) with Task Decoupling Mechanism:				
†✓	†✓	†✓	54.6	88.6
✓	†✓	†✓	**55.7**	**90.2**
✓	✓	†✓	55.0	89.0

4.3.1. *Analysis of the different components*

To assess the impact of different components, we conduct several ablation studies. Rows 2–4 in Table 3 show that SAF offers slightly less improvement compared with MGMC. However, when MGMC and SAF are combined (mAP ↑ 2.2% and top-1 ↑ 1.1%), they more effectively capture and fuse multiscale information. Rows 5 and 6 indicate that using ASA (w/GS) yields a more significant improvement than using fixed shift parameters (mAP ↑ 0.9% and top-1 ↑ 0.6%). This suggests that the dynamic shift parameter generation with Gumbel Softmax is more effective in handling occlusion issues. The final two rows show that TDM (w/$\mathcal{L}_{ort}$) further enhances performance compared with TDM without $\mathcal{L}_{ort}$ (mAP ↑ 0.5% and top-1 ↑ 0.6%), highlighting the crucial role of orthogonal loss in addressing conflicts between detection and ReID tasks.

4.3.2. *Analysis of the contributions of different stages*

We additionally analyze the contributions of our method when the SAT is applied at different stages on the PRW dataset. As depicted in Table 4 (a), incorporating the SAT in stage-1 enhances performance by 1.2% in mAP and 0.2% in top-1 compared with the baseline. Expanding its implementation to all three stages results in a 54.8% mAP and 89.3% top-1 score. This showcases the effectiveness of our SAT. In addition, we hypothesize that SAT may exhibit a synergistic effect across different stages, potentially enhancing overall performance even when its impact is not individually the most significant in some stages.

4.3.3. *Analysis of the effectiveness of decoupling*

As shown in Table 4 (b), decoupling in the first stage leads to a worse performance (mAP of 54.6% vs. 55.7%). This indicates that learning simultaneous representations for person detection and ReID is extremely challenging in the initial stage. Therefore, we remove the TDM in the first stage of our method. Removing the decoupling module in the second stage results in a decrease of 0.7% in mAP and 0.2% in top-1 score. This suggests that in the second stage, the model benefits from the decoupling module, which reduces competition between detection and ReID tasks and enhances overall performance.

To more intuitively validate the effect of orthogonality loss, we compare the correlation between detection and ReID features with and without orthogonality loss. When the cosine similarity between feature vectors approaches zero, it indicates good orthogonality between them. As shown in Fig. 7(a), without orthogonality loss, the correlation between subtask features is high, limiting the independence of the features. By contrast, in Fig. 7(b), the orthogonality loss effectively increases feature divergence, indicating that each task learns independent features. Overall, these results are consistent with the performance improvements reported in Table 3.

4.3.4. *Analysis of the context aggregator*

We evaluate the performance of the proposed context aggregator on the PRW dataset. As shown in Table 5, the first row represents the baseline method using TDM (excluding the first stage), with Res5 in all

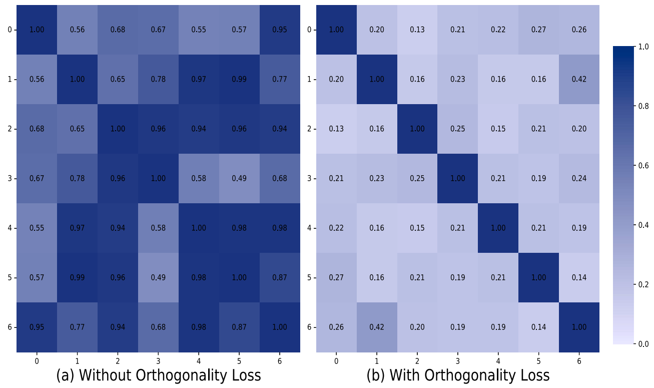


Fig. 7. The feature correlation between detection and ReID features is shown. (a) shows the correlation without orthogonality loss, and (b) shows the correlation with orthogonality loss.

Table 5
Comparison of performance of various context aggregators on the PRW dataset.

Res5	Conv Encoder	Transformer	mAP	top-1
✓			51.8	87.4
	✓		53.3	89.1
		✓	54.5	88.6
	✓	✓	55.7	90.2

Table 6
A comparison between our adaptive shift and other augmentation mechanisms. “Tokens” and “Feats” denote augmentation at the token level and feature level, respectively.

Method	Tokens	Feats	mAP	top-1
Vanilla			54.5	89.5
Jigsaw [29]	✓		53.6	88.2
Occluded Attention [23]	✓		55.1	89.3
Cutout [59]		✓	54.8	88.8
Mixup [60]		✓	54.4	88.7
Group Shift [24]		✓	54.9	89.0
Adaptive Shift (Ours)		✓	55.7	90.2

stages. Substituting Res5 with the Conv Encoder improves the results to 53.3% mAP and 89.1% top-1. The use of our SAT instead of Res5 (without Conv Encoder) yields 54.5% mAP and 88.6% top-1. Finally, combining the Conv Encoder and transformer as the hybrid context aggregator yields the top performance of 55.7% mAP and 90.2% top-1, showing the benefits of both convolution and transformer.

4.3.5. Comparison with other augmentation methods

To validate the advantage of the adaptive shift in handling occlusions, we apply recently devised strategies, such as jigsaw [29], occluded attention [23], and group shift [24], as well as previous pedestrian enhancement strategies, such as cutout and mixup, to our method. The vanilla variant of our method excludes ASA, serving as a baseline for comparison. Table 6 shows that jigsaw shows reduced performance because of token-level augmentation, likely because shuffling tokens on a small 14 × 14 feature map blurs the model’s capability to effectively search for pedestrians. Occluded attention, which considers information between different instances in a mini-batch, demonstrates improved performance to some extent. Cutout slightly increases mAP by 0.3%, and mixup does not contribute to any noticeable enhancement in accuracy. Group shift’s performance improvement is less significant than that of our ASA, which dynamically adjusts feature map positions based on occlusions, thereby enabling more accurate feature representations.

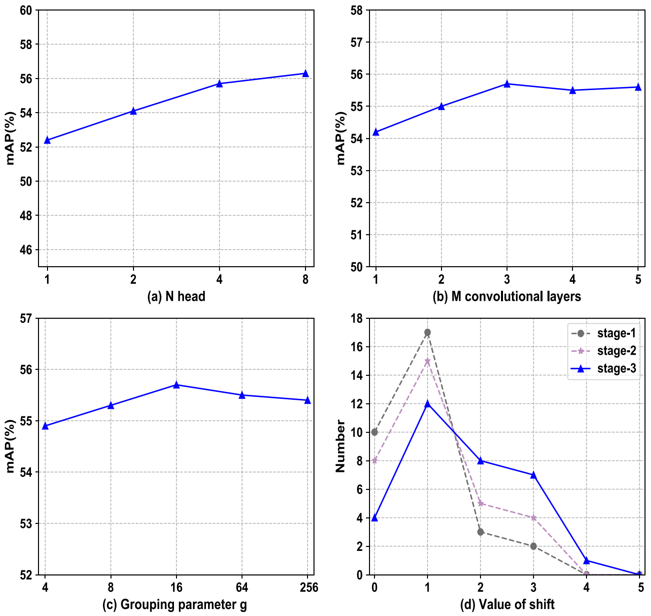


Fig. 8. Parameter analysis of our framework with different values of (a) N head, (b) M convolutional layers, and (c) grouping parameter g. The mAP on the PRW dataset is shown. (d) Analysis of shift value distribution across stages. Zoom in for better viewing.

Table 7
Validation of the generalizability of our method on different baselines. “SAT” denotes the scale-aware transformer and “TDM” denotes the task decoupling mechanism.

Baseline	SAT	TDM	CUHK-SYSU		PRW	
			mAP	top-1	mAP	top-1
NAE [22]			91.5	92.4	43.3	80.9
	✓		92.3	93.5	45.6	83.5
		✓	91.8	92.9	44.8	82.7
	✓	✓	93.4	93.8	46.3	84.2
SeqNet [32]			94.8	95.7	47.6	87.6
	✓		95.4	95.9	51.4	88.2
		✓	95.1	95.4	50.8	87.9
	✓	✓	95.6	96.1	52.4	88.7

4.3.6. Generalizability of our method

Table 7 demonstrates the generalizability of our proposed method across different baseline models, particularly NAE and SeqNet, which are widely used in person search tasks. When combined with our SAT and TDM, both baselines show marked improvements in mAP and top-1 accuracy on the CUHK-SYSU and PRW datasets. In particular, the introduction of SAT allows our model to better handle scale variations and occlusion issues in person search, and TDM effectively reduces interference between the detection and ReID tasks, enabling each subtask to independently learn its specific features. These results highlight the effectiveness and generalizability of our approach in enhancing performance across various person search models.

4.4. Parameter analysis

N head. As depicted in Fig. 8(a), raising the count of heads (N) improves the model’s accuracy (mAP) because of an expanded receptive field and enhanced multiscale information acquisition. Although N = 8 achieves the best performance, it also doubles the computational complexity compared with N = 4. To strike the best balance between performance and efficiency, we set N to 4.

M convolutional layers. Fig. 8(b) shows the effect of different convolutional layer counts on performance. Setting M to 3 yields the best result because consecutive convolutional layers enhance the perceptual

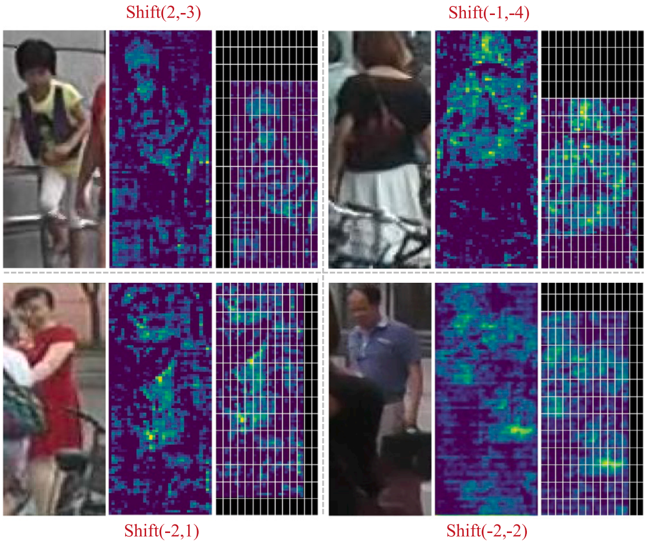


Fig. 9. Visualizing model adaptation to occlusion with adaptive shift augmentation, we observe that our model often uses larger shift values and proactively moves downward to avoid occlusions.

capabilities of these task-invariant features. Further, increasing M does not significantly improve performance and increases model complexity.

Grouping parameter g. Fig. 8(c) illustrates the impact of the grouping parameter g on performance. A larger g groups more information, and after the shift operation, each group obtains visual information from adjacent groups, increasing the model's expressive power. However, an excessively large g can cause overfitting and can reduce mAP.

Analysis of shift values. Fig. 8(d) shows the shift values learned adaptively at each stage. Each group generates two shift values, with a total of 32 values when g = 16 and with negative values unified to positive values. We observe that deeper stages tend to use larger shift values, indicating that they help avoid occlusions and improve target recognition.

4.5. *Visualization of adaptive shift augmentation*

We showcase our model's adaptation to occluded scenes by applying various shift values during image processing. Using adaptive shift operations on RoI-Align [61–65] cropped images, we compare them with the input images to demonstrate the model's transformation capabilities [66,67]. The shifts denoted as shift(x, y) provide a clear representation of direction and magnitude, enhancing the model's intuitive processing. In Fig. 9, we observe our model's preference for larger shift values in occluded scenarios. This strategic shifting aims to enhance pedestrian recognition by avoiding occluded regions. Notably, the model tends to use negative values along the y-axis, which aligns with the common occurrence of occlusions at lower body levels. These findings enhance our understanding of the adaptive shift strategy in handling complex scenes.

4.6. *Qualitative performance*

We conduct a qualitative analysis of the PRW dataset, comparing our method with COAT [23] and SeqNet [32]. In Fig. 10(a), the small size of the query target causes SeqNet to confuse it with visually similar individuals, identifying only the most obvious match. However, COAT effectively matches most targets. Enhanced by our SAM, our method outperforms both, excelling in handling scale variations. In Fig. 10(b), occlusions lead SeqNet to misidentify the target, failing to achieve a rank-1 result, and COAT tends to fixate on a distinct pink bag, resulting in incorrect outcomes. By leveraging our ASA, our method successfully focuses on unoccluded information, demonstrating robust performance despite occlusions. However, as shown in Fig. 10(c), small targets in similar clothing can still lead to errors because of the difficulty in distinguishing fine-grained features. In real-world applications, person search remains complex because of occlusions, lighting changes, and scale/pose variations.

5. Conclusion

In this study, we introduce ASTD, a novel one-step person search framework designed to extract highly discriminative features, thereby enhancing search performance and robustness. In particular, we use the SAT to address scale/pose variations and occlusions. The SAT incorporates SAM to capture and fuse multiscale information, alongside ASA, which dynamically learns shift parameters to bolster robustness against occlusions. Furthermore, the TDM ensures subtask independence and orthogonality, effectively addressing conflicts between them. The experimental results obtained on the CUHK-SYSU and PRW datasets indicate that ASTD outperforms current state-of-the-art methods.

Limitations and Future Work: Although ASTD achieves commendable results, there are areas for improvement. (1) Previous research overlooked the connections between the scene and pedestrians, considering the background as harmful. However, this information is crucial for person search, such as relationships with adjacent pedestrians or background objects. Using context-aware technologies or attention mechanisms can enhance feature discrimination. (2) Although ASTD excels at extracting discriminative features, it does not fully optimize the overall network architecture, leading to somewhat slower inference speeds compared with some existing methods. Investigating knowledge distillation to reduce network size and enhance inference speed is a promising direction for future research.

CRediT authorship contribution statement

Qixian Zhang: Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Duoqian Miao:** Supervision, Resources, Funding acquisition. **Qi Zhang:** Writing – original draft, Methodology, Conceptualization. **Changwei Wang:** Writing – original draft, Visualization. **Yanping Li:** Visualization, Formal analysis. **Hongyun Zhang:** Supervision, Funding acquisition. **Cairong Zhao:** Visualization, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The work is supported by the National Key Research and Development Program, China (Grant No. 2022YFB3104700), the National Natural Science Foundation of China (Grant Nos. 62376198, 62006172, 62076182, 62163016), the Jiangxi Double Thousand Plan, China (No. jxsq2019102088), the Jiangxi Provincial Natural Science Fund, China (No. 20212ACB202001).

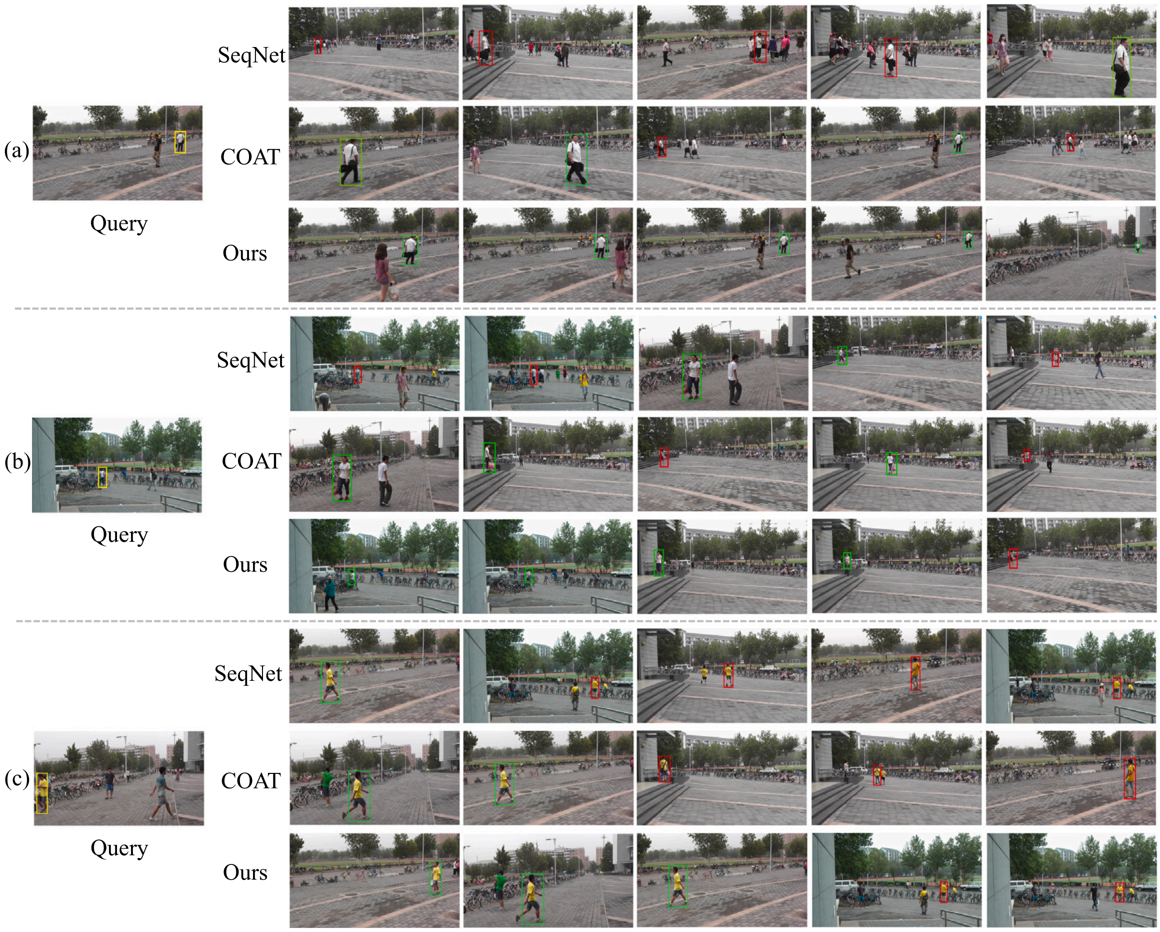


Fig. 10. Qualitative search results on the PRW dataset. (a) Comparative analysis under scale/pose variations, (b) comparative analysis under occlusion interference, and (c) error matching analysis under similar clothing. For a query person (yellow box), we display the top rank-5 search results, with red/green boxes indicating incorrect/correct results.

References

[1] Y. Xu, B. Ma, R. Huang, L. Lin, Person search in a scene by jointly modeling people commonness and person uniqueness, in: Proc. 22nd ACM Int. Conf. Multimedia, 2014, pp. 937–940.

[2] T. Xiao, S. Li, B. Wang, L. Lin, X. Wang, Joint detection and identification feature learning for person search, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit, (CVPR), 2017, pp. 3415–3424, Link.

[3] L. Zheng, S. Sun, M. Chandraker, Y. Yang, Q. Tian, Person re-identification in the wild, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit, (CVPR), 2017, pp. 1367–1376, Link.

[4] J. Zhou, B. Huang, W. Fan, Z. Cheng, Z. Zhao, W. Zhang, Text-based person search via local-relational-global fine grained alignment, Knowl.-Based Syst. 262 (2023) 110253.

[5] P. Zhang, X. Yu, X. Bai, C. Wang, J. Zheng, X. Ning, Joint discriminative representation learning for end-to-end person search, Pattern Recognit. 147 (2024) 110053.

[6] S. Hou, C. Zhao, Z. Wei, D. Miao, Improved instance discrimination and feature compactness for end-to-end person search, IEEE Trans. Circuits Syst. Video Technol. 32 (4) (2022) 2079–2090.

[7] Y. Tang, B. Li, M. Liu, B. Chen, Y. Wang, W. Ouyang, Autopedestrian: An automatic data augmentation and loss function search scheme for pedestrian detection, IEEE Trans. Image Process. 30 (2021) 8483–8496.

[8] C. Ma, L. Zhuo, J. Li, Y. Zhang, J. Zhang, Cascade transformer decoder based occluded pedestrian detection with dynamic deformable convolution and gaussian projection channel attention mechanism, IEEE Trans. Multimedia. 25 (2023) 1529–1537.

[9] S. Chan, W. Meng, J. Hu, S. Chen, Diverse-feature collaborative progressive learning for visible-infrared person re-identification, IEEE Trans. Ind. Inform. 20 (5) (2024) 7754–7763.

[10] Y. Li, Y. Liu, C. Zhao, Z. Wei, D. Miao, Occlusion-aware transformer with second-order attention for person re-identification, IEEE Trans. Image Process. 33 (2024) 3200–3211.

[11] W. Zhao, Y. Huang, G. Wang, B. Zhang, Y. Gao, Y. Liu, Multi-scale spatio-temporal feature adaptive aggregation for video-based person re-identification, Knowl.-Based Syst. 299 (2024) 111980.

[12] Z. Pang, C. Wang, H. Pan, L. Zhao, J. Wang, M. Guo, MIMR: Modality-invariance modeling and refinement for unsupervised visible-infrared person re-identification, Knowl.-Based Syst. 285 (2024) 111350.

[13] Z. Yu, P. Tiwari, L. Hou, L. Li, W. Li, L. Jiang, X. Ning, Mv-reid: 3d multi-view transformation network for occluded person re-identification, Knowl.-Based Syst. 283 (2024) 111200.

[14] M. Zhang, M. Xin, C. Gao, X. Wang, S. Zhang, Attention-aware scoring learning for person re-identification, Knowl.-Based Syst. 203 (2020) 106154.

[15] C. Zhao, X. Lv, Z. Zhang, W. Zuo, J. Wu, D. Miao, Deep fusion feature representation learning with hard mining center-triplet loss for person re-identification, IEEE Trans. Multimedia. 22 (12) (2020) 3180–3195.

[16] Y. Liu, W. Zhou, Y. Li, S. Zhao, RoSe: Rotation-invariant sequence-aware consensus for robust correspondence pruning, in: ACM Int. Conf. Multimedia, 2024.

[17] D. Chen, S. Zhang, W. Ouyang, J. Yang, Y. Tai, Person search via a mask-guided two-stream cnn model, in: Proc. Eur. Conf. Comput. Vis, (ECCV), 2018, pp. 734–750, Link.

[18] C. Han, J. Ye, Y. Zhong, X. Tan, C. Zhang, C. Gao, N. Sang, Re-id driven localization refinement for person search, in: Proc. IEEE/CVF Int. Conf. Comput. Vis, (ICCV), 2019, pp. 9814–9823, Link.

[19] C. Wang, B. Ma, H. Chang, X. Chen, TCTS: A task-consistent two-stage framework for person search, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, 2020, pp. 11952–11961, Link.

[20] H. Yao, C. Xu, Joint person objectness and repulsion for person search, IEEE Trans. Image Process. 30 (2021) 685–696.

[21] Y. Zhong, X. Wang, S. Zhang, Robust partial matching for person search in the wild, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, (CVPR), 2020, pp. 6827–6835, Link.

[22] D. Chen, S. Zhang, J. Yang, B. Schiele, Norm-aware embedding for efficient person search, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, (CVPR), 2020, pp. 12615–12624, Link.

[23] R. Yu, D. Du, R. LaLonde, D. Davila, C. Funk, A. Hoogs, B. Clipp, Cascade transformers for end-to-end person search, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, (CVPR), 2022, pp. 7267–7276, Link.

[24] M. Fiaz, H. Cholakkal, R.M. Anwer, F.S. Khan, SAT: scale-augmented transformer for person search, in: Proc. IEEE Winter Conf. Appl. Comput. Vis, (WACV), 2023, pp. 4820–4829, Link.

[25] Y. Yan, J. Li, J. Qin, S. Bai, S. Liao, F. Zhu, L. Shao, Anchor-free person search, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, (CVPR), 2021, pp. 7690–7699, Link.

[26] Z. Tian, C. Shen, H. Chen, T. He, FCOS: A simple and strong anchor-free object detector, IEEE Trans. Pattern Anal. Mach. Intell. 44 (4) (2022) 1922–1933.

[27] T. Wang, H. Liu, T. Guo, W. Shi, Pose-guided feature disentangling for occluded person re-identification based on transformer, in: Proc. AAAI Conf. Artif. Intell, vol. 36, 2022, pp. 2540–2549.

[28] G. Wang, H. Liu, Z. Wang, Y. Yang, E. Zhou, J. Sun, High-order information matters: Learning relation and topology for occluded person re-identification, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, (CVPR), 2020, pp. 6449–6458, Link.

[29] S. He, H. Luo, P. Wang, F. Wang, H. Li, W. Jiang, TransReID: Transformer-based object re-identification, in: Proc. IEEE/CVF Int. Conf. Comput. Vis, (ICCV), 2021, pp. 15013–15022, Link.

[30] C. Han, Z. Zheng, K. Su, D. Yu, Z. Yuan, N. Sang, Y. Yang, DMRNet++: Learning discriminative features with decoupled networks and enriched pairs for one-step person search, IEEE Trans. Pattern Anal. Mach. Intell. 45 (6) (2023) 7319–7337.

[31] C. Han, Z. Zheng, C. Gao, N. Sang, Y. Yang, Decoupled and memory-reinforced networks: Towards effective feature learning for one-step person search, in: Proc. AAAI Conf. Artif. Intell, vol. 35, 2021, pp. 1505–1512.

[32] Z. Li, D. Miao, Sequential end-to-end network for efficient person search, in: Proc. AAAI Conf. Artif. Intell, (3) 2021, pp. 2011–2019.

[33] Z. Cai, N. Vasconcelos, Cascade R-CNN: Delving into high quality object detection, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, (CVPR), 2018, pp. 6154–6162, Link.

[34] X. Lan, X. Zhu, S. Gong, Person search by multi-scale matching, in: Proc. Eur. Conf. Comput. Vis, (ECCV), 2018, pp. 536–552, Link.

[35] X. Chang, P.-Y. Huang, Y.-D. Shen, Y. Yang, A.G. Hauptmann, RCAA: Relational context-aware agents for person search, in: Proc. Eur. Conf. Comput. Vis, (ECCV), 2018, pp. 84–100, Link.

[36] W. Dong, Z. Zhang, C. Song, T. Tan, Bi-directional interaction network for person search, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, (CVPR), 2020, pp. 2839–2848, Link.

[37] B. Munjal, S. Amin, F. Tombari, F. Galasso, Query-guided end-to-end person search, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, (CVPR), 2019, pp. 811–820, Link.

[38] L. Jaffe, A. Zakhor, Gallery filter network for person search, in: Proc. IEEE Winter Conf. Appl. Comput. Vis, (WACV), 2023, pp. 1684–1693, Link.

[39] Z. Song, C. Zhao, G. Hu, D. Miao, Learning scene-pedestrian graph for end-to-end person search, IEEE Trans. Ind. Inform. 20 (2) (2024) 2979–2990.

[40] M. Cao, Y. Bai, Z. Zeng, M. Ye, M. Zhang, An empirical study of clip for text-based person search, in: Proc. AAAI Conf. Artif. Intell, vol. 38, 2024, pp. 465–473.

[41] Z. Song, G. Hu, C. Zhao, Diverse person: Customize your own dataset for text-based person search, in: Proc. AAAI Conf. Artif. Intell, vol. 38, 2024, pp. 4943–4951.

[42] S. Dou, C. Zhao, S. Zhang, W.-S. Zheng, W. Zuo, Human co-parsing guided alignment for occluded person re-identification, IEEE Trans. Image Process. 32 (2023) 458–470.

[43] Y. Li, J. He, T. Zhang, X. Liu, F. Wu, Diverse part discovery: Occluded person re-identification with part-aware transformer, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, (CVPR), 2021, pp. 2898–2907, Link.

[44] Q. Zhang, J. Wu, D. Miao, C. Zhao, Q. Zhang, Attentive multi-granularity perception network for person search, Inform. Sci. 681 (2024) 121191.

[45] A. Dosovitskiy, L. Beyer, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: Proc. Int. Conf. Learn. Representations, 2021, Link.

[46] H. Wang, J. Shen, Y. Liu, Y. Gao, E. Gavves, Nformer: Robust person re-identification with neighbor transformer, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, (CVPR), 2022, pp. 7297–7307, Link.

[47] Y. Li, D. Miao, H. Zhang, J. Zhou, C. Zhao, Multi-granularity cross transformer network for person re-identification, Pattern Recognit. 150 (2024) 110362.

[48] Y. Liu, B.N. Zhao, S. Zhao, L. Zhang, Progressive motion coherence for remote sensing image matching, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–13.

[49] Y. Liu, Y. Li, L. Dai, C. Yang, L. Wei, T. Lai, R. Chen, Robust feature matching via advanced neighborhood topology consensus, Neurocomputing 421 (2021) 273–284.

[50] J. Cao, Y. Pang, R.M. Anwer, H. Cholakkal, J. Xie, F.S. Khan, PSTR: End-to-end one-step person search with transformers, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, (CVPR), 2022, pp. 9458–9467, Link.

[51] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: Proc. Eur. Conf. Comput. Vis, (ECCV), Springer, 2020, pp. 213–229, Link.

[52] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: Proc. IEEE Int. Conf. Comput. Vis, (ICCV), 2017, pp. 2961–2969, Link.

[53] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, Y. Li, Maxvit: Multi-axis vision transformer, in: Proc. Eur. Conf. Comput. Vis, (ECCV), Springer, 2022, pp. 459–479.

[54] E. Jang, S. Gu, B. Poole, Categorical reparameterization with gumbel-softmax, in: Proc. Int. Conf. Learn. Representations, 2017, Link.

[55] W. Dong, Z. Zhang, C. Song, T. Tan, Instance guided proposal network for person search, in: Proc. IEEE/CVF Conf. Comput Vis. Pattern Recognit, (CVPR), 2020, pp. 2585–2594, Link.

[56] Y. Yan, Q. Zhang, B. Ni, M. Xu, X. Yang, Learning context graph for person search, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, (CVPR), 2019, pp. 2158–2167, Link.

[57] C. Zhao, Z. Chen, S. Dou, Z. Qu, J. Yao, J. Wu, D. Miao, Context-aware feature learning for noise robust person search, IEEE Trans. Circuits Syst. Video Technol. 32 (10) (2022) 7047–7060.

[58] W. Chen, X. Xu, J. Jia, Y. Wang, F. Wang, X. Sun, Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, (CVPR), 2023, pp. 15050–15061, Link.

[59] T. DeVries, G.W. Taylor, Improved regularization of convolutional neural networks with cutout, 2017, arXiv preprint.Link.

[60] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, Mixup: Beyond empirical risk minimization, in: Proc. Int. Conf. Learn. Representations, 2018, Link.

[61] Q. Zhang, L. Cao, C. Shi, Z. Niu, Neural time-aware sequential recommendation by jointly modeling preference dynamics and explicit feature couplings, IEEE Trans. Neural Netw. Learn. Syst. 33 (10) (2021) 5125–5137.

[62] K. Yuan, D. Miao, Y. Yao, H. Zhang, X. Zhao, Feature selection using zentropy-based uncertainty measure, IEEE Trans. Fuzzy Syst. 32 (4) (2024) 2246–2260.

[63] K. Yi, Q. Zhang, W. Fan, P. Wang, D. Lian, L. Cao, Z. Niu, Frequency-domain mlps are more effective learners in time series forecasting, in: Proc. Adv. Neural Inf. Process. Syst, vol. 36, 2024, Link.

[64] Z. Gong, Q. Zhang, G. Bao, L. Zhu, Y. Zhang, L. Hu, D. Miao, Lite-mind: Towards efficient and robust brain representation learning, in: ACM Int. Conf. Multimedia, 2024.

[65] Y. Zhang, Y. Liu, D. Miao, Q. Zhang, Y. Shi, L. Hu, MG-ViT: a multi-granularity method for compact and efficient vision transformers, in: Proc. Adv. Neural Inf. Process. Syst, vol. 36, 2024, Link.

[66] K. Yuan, D. Miao, W. Pedrycz, W. Ding, H. Zhang, Ze-HFS: Zentropy-based uncertainty measure for heterogeneous feature selection and knowledge discovery, IEEE Trans. Knowl. Data En. (2024).

[67] D. Guo, W. Xu, Y. Qian, W. Ding, M-FCCL: Memory-based concept-cognitive learning for dynamic fuzzy data classification and knowledge fusion, Inform. Fusion. 100 (2023) 101962.