Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.067264

ARTICLE

ConvNeXt-Driven Dynamic Unified Network with Adaptive Feature Calibration for End-to-End Person Search

Xiuchuan Cheng[1], Meiling Wu[1], Xu Feng[1], Zhiguo Wang[2], Guisong Liu[2] and Ye Li[2,*]

[1]Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen, 518000, China
[2]School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, 610054, China

*Corresponding Author: Ye Li. Email: liyeuestc@uestc.edu.cn

Received: 28 April 2025; Accepted: 15 July 2025; Published: 23 September 2025

ABSTRACT: The requirement for precise detection and recognition of target pedestrians in unprocessed real-world imagery drives the formulation of person search as an integrated technological framework that unifies pedestrian detection and person re-identification (Re-ID). However, the inherent discrepancy between the optimization objectives of coarse-grained localization in pedestrian detection and fine-grained discriminative learning in Re-ID, combined with the substantial performance degradation of Re-ID during joint training caused by the Faster R-CNN-based branch, collectively constitutes a critical bottleneck for person search. In this work, we propose a cascaded person search model (SeqXt) based on SeqNet and ConvNeXt that adopts a sequential end-to-end network as its core architecture, artfully integrates the design logic of the two-step method and one-step method framework, and concurrently incorporates the two-step method's advantage in efficient subtask handling while preserving the one-step method's efficiency in end-to-end training. Firstly, we utilize ConvNeXt-Base as the feature extraction module, which incorporates part of the design concept of Transformer, enhances the consideration of global context information, and boosts feature discrimination through an implicit self-attention mechanism. Secondly, we introduce prototype-guided normalization for calibrating the feature distribution, which leverages the archetype features of individual identities to calibrate the feature distribution and thereby prevents features from being overly inclined towards frequently occurring IDs, notably improving the intra-class compactness and inter-class separability of person identities. Finally, we put forward an innovative loss function named the Dynamic Online Instance Matching Loss Function (DOIM), which employs the hard sample assistant method to adaptively update the lookup table (LUT) and the circular queue (CQ) and aims to further enhance the distinctiveness of features between classes. Experimental results on the public datasets CUHK-SYSU and PRW and the private dataset UESTC-PS show that the proposed method achieves state-of-the-art results.

KEYWORDS: Person search; Re-ID; SeqNet; ConvNeXt

1 Introduction

Person search has emerged as a frontier research domain in computer vision, demonstrating significant utility in public security and social governance applications such as criminal investigation support, cross-terminal tracking, and emergency response. This technology seeks to accurately localize and identify target individuals from large-scale surveillance video databases using query person images, relying on two interdependent sub-tasks: pedestrian detection and person re-identification (Re-ID). Pedestrian detection, the foundational step, employs advanced computer vision algorithms to extract human instances from complex scenes by precisely delineating bounding boxes, focusing exclusively on spatial localization without

identity discrimination. In contrast, Re-ID operates on cropped pedestrian regions to discriminatively match query identities across candidate bounding boxes, addressing the cross-view identity correspondence problem. Although individually critical for video surveillance, these tasks exhibit inherent functional boundaries: detection prioritizes coarse-grained spatial localization, while Re-ID demands fine-grained feature discrimination. Their complementary yet conflicting optimization objectives, first identified by Xu et al. [1], motivate the integration of these sub-tasks into a unified framework. Person search remains a challenging problem, confronting technical hurdles such as illumination variation, viewpoint changes, occlusions, pose diversity, and low-resolution imagery, while requiring the joint optimization of detection-localization and Re-ID-discrimination in a coherent architecture.

Person search methodologies are principally categorized into two paradigms: two-step and one-step approaches. The former employs distinct pedestrian detection and recognition modules, using detectors like Faster Region-based Convolutional Neural Networks (R-CNN) to generate bounding boxes, crop the detected regions, and feed them into a dedicated recognition network; this approach yields high accuracy but entails significant computational costs. The latter integrates detection and recognition within a unified framework, leveraging a shared feature backbone to directly predict pedestrian bounding boxes and discriminative recognition features from raw images to enable end-to-end training.

In recent years, person search techniques have increasingly embraced a one-step framework typified by OIMNet [2], which integrates pedestrian detection and feature extraction into a unified architecture by leveraging Faster R-CNN [3] to enhance multi-scale feature capture and employing L2 normalization alongside an online instance matching loss function to optimize discriminative feature representation, thereby effectively enhancing the accuracy and generalization capability of person re-identification in complex visual environments.

Despite significant progress in OIMNet-based person search technology, two major challenges persist: first, under the assumption of a random distribution of pedestrian characteristics in Fig. 1a, the effectiveness of L2 normalization is compromised by feature zero-centrality and inter-channel variance inconsistency, which in practice leads to diminished feature discriminability as demonstrated in Fig. 1b; second, Batch Normalization fails to function effectively owing to class imbalance and the diversity of person IDs in datasets, inducing feature distributions biased toward common identities that degrade discriminative power in Fig. 1c. Additionally, the fixed momentum update strategy of the current Online Instance Matching (OIM) loss function may attenuate target prototype features and reduce inter-class discriminability when addressing cross-class confusion samples. Thus, enhancing the discriminative ability of person features and improving search performance critically hinges on refining the feature distribution processing method and optimizing the prototype update strategy within the OIM loss framework.

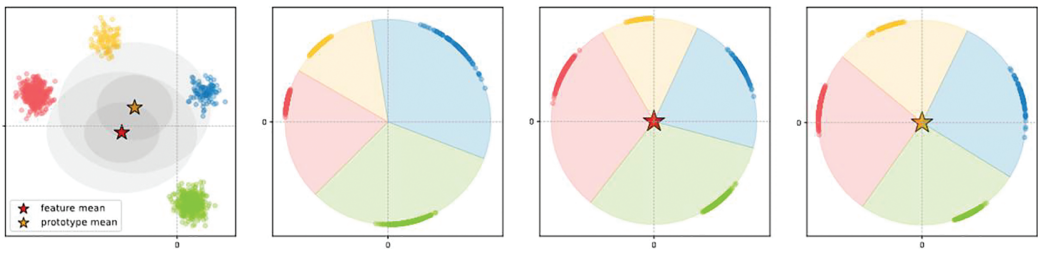


Figure 1: Different normalization methods. Different colors represent different person ids, and red and yellow asterisks represent the mean values obtained from the input features and ID prototypes, respectively

This paper introduces SeqXt, an innovative cascading person search model built upon SeqNet's core architecture that seamlessly integrates the hierarchical task processing advantages of two-step frameworks with the end-to-end training efficiency of one-step methodologies, forming a coherent composite architecture that balances logical modularity and computational effectiveness. SeqXt inherits the two-step framework's structured pipeline for sequential pedestrian detection and re-identification while adopting the one-step framework's seamless integration design, enabling systematic subtask processing without compromising training efficiency. Optimizations to the baseline network occur at two primary levels: first, the introduction of Prototype-guided Norm (PN), a novel normalization layer that leverages individual identity prototype features to calibrate feature distributions on the hypersphere, thereby suppressing over-reliance on high-frequency person IDs and enhancing intra-class compactness and inter-class separability as illustrated in Fig. 1d; second, the development of Dynamic Online Instance Matching Loss (DOIM), which employs a hard-sample auxiliary strategy to adaptively update the Look-up Table (LUT) and Circular Queue (CQ), aiming to deepen inter-class feature discrimination and improve overall recognition accuracy. Additionally, ConvNeXt-Base is selected as the feature extraction backbone, incorporating Transformer-inspired design elements to enhance global context modeling through implicit self-attention mechanisms, thereby boosting the discriminative power of output features via deep integration of multi-scale visual context.

In summary, the core contributions of this paper are summarized as follows.

1. In this paper, we use the ConvNeXt-Base architecture as a person feature extraction network, aiming to refine a more discriminative feature representation. To further balance the class distribution in the dataset and enhance the performance of person ID in terms of intra-class compactness and inter-class discrimination, this paper introduces the prototypical guided normalization method to calibrate the feature distribution.
2. This paper also proposes a dynamic online instance matching method. By constructing a loss function based on the prototype feature update principle assisted by difficult samples, the feature diversity between classes is improved to guide the model to learn effectively. Experimental results on the private dataset UESTC-PS show that the proposed method outperforms the existing public models.

2 Related Work

2.1 CNN-Based Person Search

Person search is a hot topic in computer vision. According to the training mechanism, person search is divided into two-step method and one-step method. In the two-step method, the person's location is initially determined by the person detector. Subsequently, the re-identification model is employed to extract features, and the search results are obtained following integration. Zheng et al. [4] explored the synergistic effect between person detection and re-identification models, and adopted cascaded fine-tuning training strategy and Confidence Weighted Similarity (CWS) matching algorithm to effectively improve person search accuracy. Lan et al. [5] focused on the resolution diversity of person search and found the importance of multi-scale matching. They used Cross-level semantic alignment (CLSAs) technology to solve the feature matching problem caused by resolution differences and ensure that the model is stable at various scales. Han et al. [6] noted that the bounding boxes generated by the independent person detection model did not fulfill the requirements for re-identification. They designed a RoI transformation module, so that the gradient information of the re-identification model was fed back to the detection model, and the bounding box generation was optimized, which was both accurate and conducive to re-identification, and improved the performance of the person search system.

The one-step person search model integrates detection and re-identification in a single framework to achieve end-to-end synchronous training. Early models are based on the Faster R-CNN architecture, add a fully connected layer to generate re-identification features, and use OIM or central loss function to supervise model training. Yan et al. [7] found that multiple persons in an image have internal correlation, and used graph convolution technology to model person interaction and improve the discrimination of target person representation. Munjal et al. [8] utilized the presence of unknown persons as prior knowledge to guide the feature representation of selected individuals. This helps establish correlations between them, enhancing matching accuracy and reducing false alarms from irrelevant person bounding boxes. Zhang et al. [9] observed that performing pooling on partial regions of feature maps helps capture richer contextual information. Chen et al. [10] proposed NAE network to decompose feature vectors into modulus length and Angle, which serve detection and re-identification respectively and alleviate the conflict of objective functions. Li and Miao [11] designed SeqNet, which regarded person detection and re-identification as a gradual process and was processed sequentially by sub-networks. The proposed network, inspired by the sequential end-to-end network SeqNet, employs a sequential framework that combines the advantages of the two-step and one-step frameworks to incrementally solve the detection and re-identification tasks.

2.2 Transformer-Based Person Search

Since the emergence of the ViT model [12] in image recognition tasks, it has rapidly penetrated into multiple application scenarios of computer vision with its excellent representation ability, especially in the field of person re-identification [13,14]. Li et al. [13] further adopted a component-aware Transformer in their research to develop a set of component detection techniques that can deal with occlusion in person re-identification, and significantly enhance the recognition effect in complex occlusion scenes by accurately locating and effectively integrating local features. Wang et al. [14] introduced the concept of neighborhood Transformer, aiming to mine the spatial relationship of adjacent features, and then construct a highly robust feature representation for person re-identification task. Zhang et al. [15] designed a feature calibration strategy within the Transformer framework, which cleverly uses the underlying features as global prior guidance to realize the fine adjustment and optimization of features in person re-identification. Chen and Xu [16] designed a specialized Re-ID Transformer, which fully fuses contextual information through self-attention mechanisms and employs multi-cross attention layers to extract fine-grained local features, fully verifying the effectiveness of its sequential framework design and Re-ID feature encoding mechanism. Song et al. [17] proposed a dedicated multi-query Transformer decoder for joint pedestrian detection and feature representation learning. By leveraging adjacent objects and multi-scale features, it effectively locates and learns target pedestrian features. Meanwhile, margin ranking loss is adopted to enhance matching of pedestrians with the same identity, facilitating cross-camera association of instances with the same identity, thus achieving joint optimization of pedestrian detection and feature learning as well as effective matching of pedestrians with the same identity. Lv et al. [18] constructed a feature extraction network based on Transformer and a context-enhanced region recognition head network, realizing end-to-end joint optimization. Feng et al. [19] replaced traditional CNNs with a dual Transformer architecture, maintaining comparable recognition accuracy while reducing model parameters. By introducing an occlusion attention mechanism, they significantly enhanced the model's ability to learn pedestrian features for small-scale targets and in occluded scenarios.

Recently, two state-of-the-art works, PSTR [20] and COAT [21], have applied Transformers to person search tasks, revealing its possibility in complex scenes and object tracking. Based on the DETR [22] framework, PSTR cleverly uses the encoder-decoder structure to achieve global information capture and local detail depiction in the target detection process, and realizes the person re-identification task execution

through the decoder link. However, COAT is based on the cascade RCNN [23] system and refines the discriminative features through multi-stage learning. It is worth mentioning that COAT adopts an explicit multi-scale convolutional Transformer in each stage, which aims to actively cope with the challenges caused by target scale changes in person search, and ensure that the model can maintain stable recognition performance under different scale conditions. In this paper, we use the ConvNeXt which integrates the characteristics of Transformer as a feature extraction tool, because it combines the advantages of Transformer in global dependency capture and self-attention mechanism, and can simulate the Transformer's ability to process large-scale context information within the CNN architecture. At the same time, we implicitly implement the self-attention mechanism, which ensures computational efficiency and improves the performance of the network to a level comparable to or even higher than Transformer. The proposed model meets the requirements of efficient integration of multi-scale information and model adaptability in the field of computer vision, and provides a more accurate and reliable feature extraction scheme for person search tasks.

3 Method

In this section, we describe our proposed SeqXt in detail. First, we outline the network architecture of SeqXt in Section 3.1. Second, the details of our redesigned dynamic online instance matching function are elaborated in Section 3.2.

3.1 SeqXt Architecture

3.1.1 Overall Architecture

For the end-to-end approach, researchers designed a multi-task framework based on Faster R-CNN and fused with Region Proposal Network (RPN) to generate candidate regions and pass these proposals to subsequent parallel branches, which are responsible for detection and re-identification tasks, respectively. However, it should be noted that the features extracted by the proposed framework originate from preliminary, low-precision candidates rather than precisely defined bounding boxes. While the impact of these non-optimal features may be relatively limited in more macroscopic classification tasks, their negative effects become particularly pronounced in fine-grained person re-identification tasks that demand extremely high accuracy, significantly diminishing the system's recognition performance. The root cause of the above problems is that the person search framework based on Faster R-CNN is designed in parallel. Due to the concurrent execution of detection and re-identification tasks at the same time, it is impossible to ensure that accurate bounding box information has been obtained before extracting person re-identification features. Conversely, although the two-stage solution circumvents this problem, the detection and Re-ID tasks are processed by two independent models sequentially. However, the drawbacks of this method include its time-consuming nature and high demand for computing resources.

In view of the above observations, the serialized end-to-end network SeqNet is chosen as the baseline model in this paper. The network skillfully combines the advantages of the two-step method and the one-step method to complete the person detection and re-identification task sequentially. As illustrated in Fig. 2, the serialized end-to-end network framework comprises two main branches: the first is the person pre-detection branch (depicted in Stage 1), responsible for generating high-quality person bounding boxes. This is followed by the person re-identification branch, depicted in Stage 2, responsible for fine-tuning the person bounding boxes and extracting features. Through this design, the baseline network not only retains the sequential processing flow of the two-step method to ensure accurate person localization for the re-identification stage, but also realizes an end-to-end training method to improve efficiency.

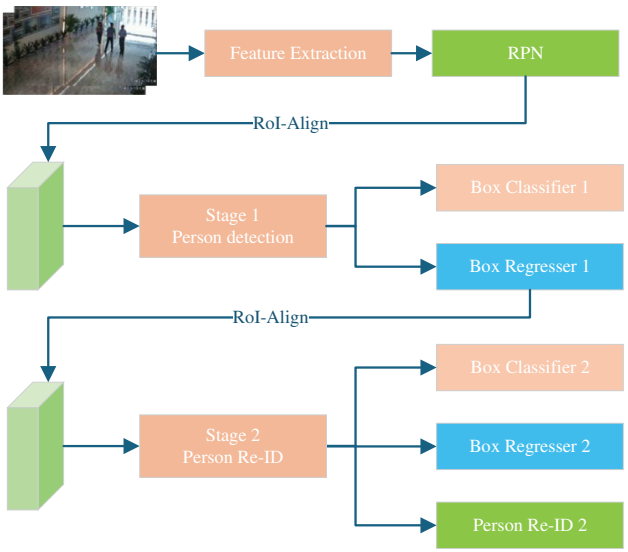


Figure 2: Overall architecture of SeqXt. The model consists of two stages. The first stage follows the standard Faster R-CNN framework to generate person bounding boxes, and the second stage adopts the standard perceptual embedding method to complete the re-identification task

While the one-step model demonstrates integration advantages in person search, its training process cannot avoid the inherent contradiction between the two core subtasks. There exists a notable conflict between the two tasks in feature learning. Person detection emphasizes universal features, focusing on accurately delineating person contours and achieving robust boundary division. Conversely, person re-identification focuses on learning individual differences, precisely identifying the identity of specific individuals, and relies on distinguishing personality traits. The optimization goal of the detection task is to compress the spatial distribution of the person feature vector, separate it from the background vector, and establish category boundaries, but it disregards the individual identity attribute of the person. Forcing the person and the background vector to be in the same dimension will limit the spatial scope of the person vector, resulting in the contraction of the Angle and the loss of discrimination, affecting the re-identification task, and it is difficult to identify the person identity, which violates the purpose of Re-ID.

Considering the aforementioned contradictions, this paper introduces an explicit vector geometric feature decomposition mechanism when constructing the second-stage person re-identification network. The aim is to cleverly balance the conflicting demands between person detection and re-identification subtasks, as illustrated in Fig. 3. Specifically, this model opts to have person detection and re-identification share the underlying feature representation. These feature vectors are then decomposed in the polar coordinate system during the final output stage. Each eigenvector is thus decomposed into two fundamental elements: the vector norm r and the angle θ. Among them, the vector norm r is given the mission of person detection, and its value is used as an effective index to measure the confidence of person bounding box detection, which helps the model to accurately locate persons in complex scenes and effectively segment them from the background. The angle θ encapsulates the crucial information for the re-identification task. By calculating the cosine similarity of the feature vectors among individuals, the relative positional relationships between persons in the feature space are quantified, facilitating accurate identification of specific person identities. This cosine similarity measure, based on the angle θ, has been extensively utilized in the field of person re-identification, with its efficiency and robustness thoroughly validated. In this way, the proposed model skillfully uniifies the mutually constrained person detection and re-identification goals into the same geometric decomposition

framework of feature vectors, realizing their harmonious coexistence and complementary advantages in feature learning and representation utilization, and providing a novel and effective solution to solve the core conflict problem in one-step person search models.

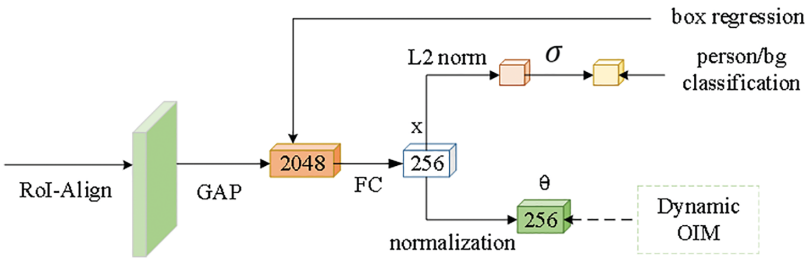


Figure 3: Re-ID network. The model applies global average pooling and a fully connected layer to obtain a 256-dimensional feature vector x, which is then placed in polar coordinates

3.1.2 Feature Extraction

To enhance the network's capability to extract and represent detailed person features, this paper upgrades the backbone component of the original network from the ResNet50 model to the advanced ConvNeXt architecture. ConvNeXt incorporates the architectural wisdom and design strategy of the Swin Transformer, renowned for its performance in object detection, thereby exhibiting improved recognition accuracy while maintaining the same level of FLOPs (Floating Point Operations per Second).

Fig. 4 provides a detailed illustration of the network structure layout of ConvNeXt. The symbol k represents the size of the convolution kernel, reflecting the sampling range of spatial features within the network. s represents the step size of the convolution kernel, determining the sampling interval of the network on the feature map. ×n represents the number of repeated stacks of a specific block type, directly impacting the network's depth and model capacity. This parameter holds decisive significance for the network's fitting ability and feature abstraction level. It's noteworthy that ConvNeXt introduces a layer scale mechanism at the end of each block, adding an extra set of learnable parameters to the output features of each layer. This allows the model to more accurately focus on key features, effectively enhancing the extraction efficiency of fine-grained person features. In addition, a depth-wise separable convolution of 7×7 was used to broaden the receptive field of the model and capture more global information. Additionally, the use of nonlinear projection operations, such as activation functions and batch normalization, is minimized to improve the transmission efficiency of features within the network and effectively enhance its expression.

According to the number of block stacks, four versions of T/S/B/L specifications are derived from ConvNeXt, whose computational complexity is exactly the same as the corresponding version in Swin Transformer. In this experiment, we choose the ConvNeXt-B model for its balanced performance and efficiency. The stacking configuration of its four blocks is (3, 3, 27, 3), and the number of output channels for each block is set to (128, 256, 512, 1024), respectively, aiming to efficiently capture and accurately express fine-grained person features.

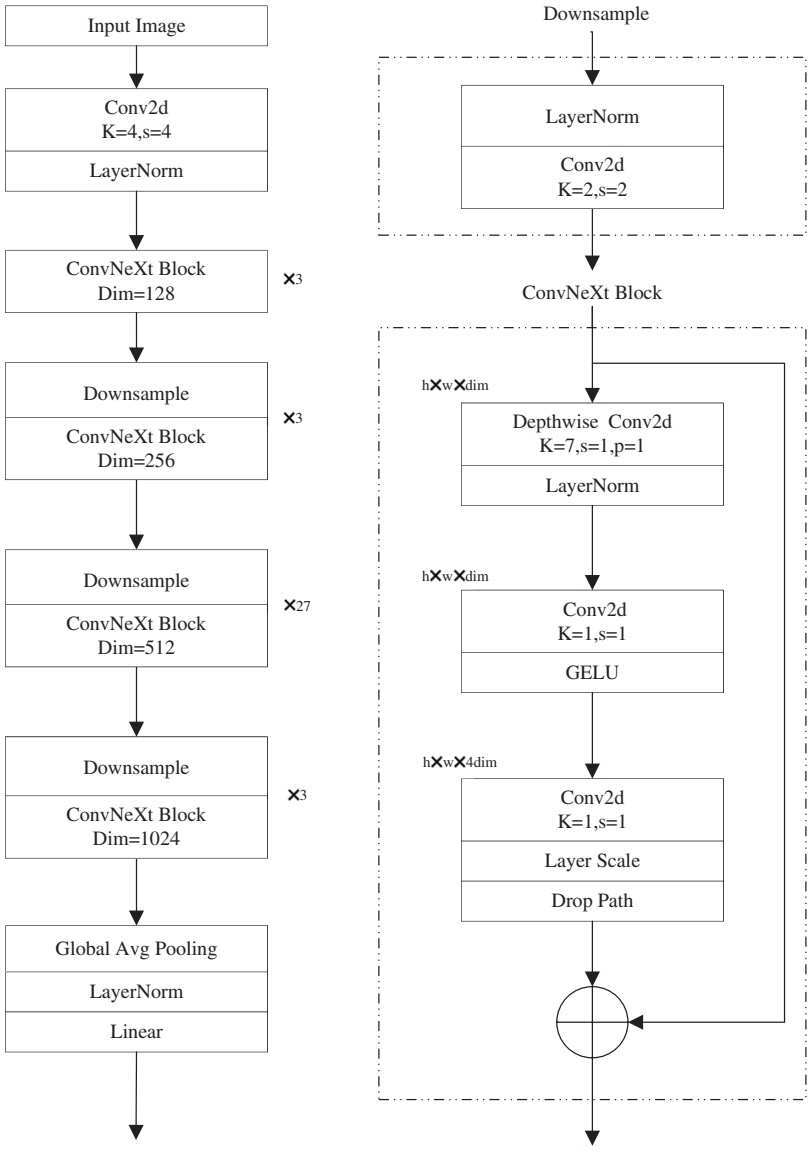


Figure 4: ConvNeXt network architecture

3.1.3 Prototype-Guided Norm

In traditional Batch Normalization (Batch Norm) technology, the system directly utilizes the channel feature statistics of the input features to correct the feature distribution. However, this operation often leads to the distribution being overly influenced by the frequency of individual ID tags, resulting in a bias towards high-frequency IDs in the feature distribution. In this paper, we propose a novel normalization method called Prototype-guided Norm (PN). Instead of directly calibrating the distribution based on input features, the PN approach advocates for using mini-batch statistics computed on prototype features corresponding to different individual IDs for adjustment. To obtain the prototype features of a specific ID, we first average the features corresponding to that ID within a small batch of data. This allows PN to adjust the feature distribution more accurately by leveraging insights into the characteristics of individual IDs, thus mitigating the dominant influence of high-frequency IDs on the overall distribution and enhancing the model's ability to

recognize various IDs in a more balanced manner. As illustrated in Fig. 5, we visually compare the differences between Prototype-guided Norm and BatchNorm in handling feature distribution, demonstrating how PN effectively addresses the limitations of BatchNorm in dealing with uneven ID tag frequencies by introducing the prototype-guided mini-batch statistics mechanism.

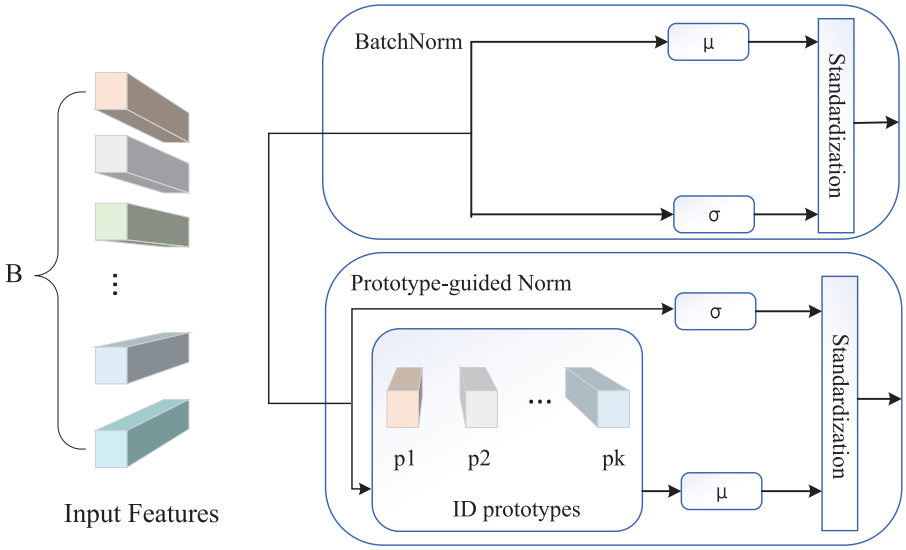


Figure 5: Difference between batch normalization and prototype bootstrapping normalization

Specifically, we denote a set of features, $X = \{x^1, \cdots, x^B\}$, where B is the size of each batch and the corresponding set of ID labels $Y = \{y^1, \cdots, y^B\}$ where $y_i \in \{1, \cdots, L\}$. We denote by $X^i(d)$ the d-th channel element x^d in the i-th feature. We first obtain a prototype feature representing the i-th ID, denoted as $p_i \in R^d$, as follows:

$$p_i(d) = \frac{\sum_{b=1}^{B} \mathrm{X}^b(d)\varphi\left[y^b = i\right]}{\sum_{b=1}^{B} \varphi\left[y^b = i\right]} \tag{1}$$

$$\varphi\left[y^b = i\right] = \begin{cases} 1, \text{ True} \\ 0, \text{ Otherwise} \end{cases} \tag{2}$$

where $\varphi[\cdot]$ is an indicator function that takes the value 1 if the argument is true and 0 otherwise. After obtaining the prototype feature p_i, we can calculate the mean and variance vectors of the input feature X, denoted by $\mu \in R^D$ and $\sigma \in R^D$, respectively, as shown in the following formula:

$$\mu(d) = \frac{1}{K}\sum_{k=1}^{K} p_i(d) \tag{3}$$

where K is the number of unique ids in the set of ID labels Y. The features are then normalized by $\frac{\mathrm{X}^b(d) - \mu(d)}{\sigma(d)}$. PN uses the weighted sum of input features instead of directly using the input features, where the weight of the t ID can be expressed as $\frac{1}{K\sum_{b=1}^{B}\varphi[y^b = i]}$, which is inversely proportional to the number of appearances of a certain ID. In other words, PN adaptively assigns larger weight values to ID features that occur less frequently in each batch, and sets smaller weight values to ids that occur more frequently. This makes the average value biased towards less frequent ids, thus increasing the difference between classes.

3.2 Dynamic OIM

3.2.1 Online Instance Matching

The online instance matching loss function is the basis of the dynamic online instance matching loss function. The main difference between person search and re-identification is that there are persons with unlabeled identities in the dataset. Xiao et al. [2] proposed Online Instance Matching (OIM) loss function, which uses unlabeled data as negative samples to increase the class diversity of person identity classifier. Specifically, there are $L + Q$ persons in the training dataset. The online instance matching loss function constructs a lookup table V to store the labeled person vector, the size of which is $L \times d$. The circular queue U is used as a memory matrix to store the unlabeled person vectors with size $Q \times d$. Together, they form the projection matrix $W \in R^{(L+Q)\times d}$. Given a person re-identification vector $x \in R^d$, the cosine similarity between the vector $x \in R^d$ and W can be directly calculated by matrix multiplication.

$$s = Wx \in R^{(L+Q)} \tag{4}$$

Given that x represents a person, the probability $P - i$ of x belonging to class i can be normalized by the softmax function:

$$p_i = \frac{\exp\left(e_i/\tau\right)}{\sum_{j=1}^{L+Q}\exp\left(e_j/\tau\right)} = \frac{\exp\left(v_i^T x_i/\tau\right)}{\sum_{j=1}^{L}\exp\left(v_j^T x_i/\tau\right) + \sum_{k=1}^{Q}\exp\left(u_k^T x_i/\tau\right)} \tag{5}$$

Among them $[v_1, v_2, \cdots, v_N] \in V$, $[u_1, u_2, \cdots, u_Q] \in U$, τ to control the temperature of the sharp degree of probability distribution coefficient. The objective function is a form that minimizes the negative log-likelihood:

$$L_{OIM} = -E_x[logp_i], i = 1, 2, \ldots, N. \tag{6}$$

During the training process, the circular queue is updated in the form of deleting old vectors and adding new vectors, and the size of the whole circular queue is guaranteed to be constant. While the lookup table is incrementalized with momentum η, the update principle of prototype feature v_i in each iteration is as follows.

$$v_i \leftarrow \eta v_i + (1-\eta)x_i, \eta \in [0,1] \tag{7}$$

3.2.2 Hard Example Auxiliary

It has been observed that the conventional OIM prototype feature update principle encounters challenges in maintaining accurate prototype feature updates, particularly when confronted with a significant number of inter-class challenging samples. In scenarios where numerous individuals exhibit highly similar appearance characteristics—such as identical clothing colors, similar body shapes, and matching perspectives—the OIM loss function may inadvertently update the prototype features of misidentified identities to align with those of existing identity prototypes. This phenomenon often leads to a notable decrease in the feature differentiation between classes.

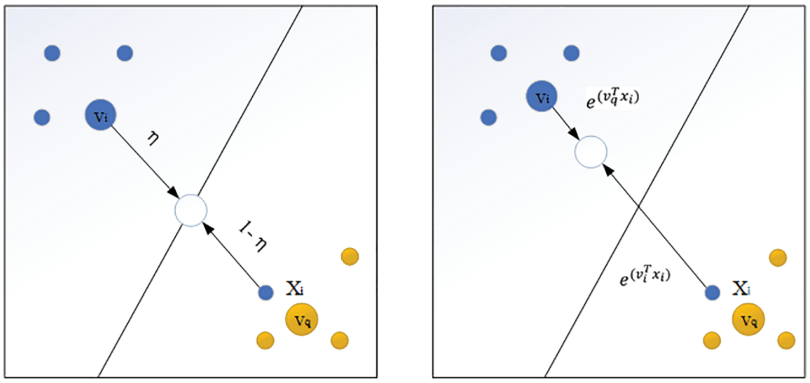
Therefore, this paper proposes a Hard Example Auxiliary (HEA) strategy. For the original weight parameter η, the auxiliary weight of difficult samples is used to replace, so as to achieve the purpose of pushing the inter-class features far away when there are more difficult samples between classes, and enhance the independent significance of the prototype features between classes. For the person x_i, whose input

belongs to class i, the corresponding features stored in the existing LUT table V are v_i, and the most difficult inter-class sample v_q of v_i is found, namely:

$$q = argmax_{p\in\{1,2,\ldots,L\}} v_p^T v_i \tag{8}$$

As shown in Fig. 6a, the difficult sample v_q is closer to x_i than the prototype v_i, that is, the similarity of $v_q - x_i$ is greater than $v_i - x_i$. When we update the prototype feature this time, if the original prototype feature update principle is adopted and η parameter value is set to 0.4, the prototype will be closer to the difficult sample and the difference between classes will be further reduced. If the corresponding weight of the original v_i is increased and the corresponding weight of x_i is decreased, the original prototype accounts for a larger part of the updated prototype features, so that the poor quality features have less impact on the prototype update, and better network training effect can be achieved, as shown in Fig. 6b. The corresponding prototype feature update principle is as follows.

$$v_i \leftarrow \frac{\exp\left(v_q^T x_i/\tau\right)}{\exp\left(v_q^T x_i/\tau\right)+\exp\left(v_i^T x_i/\tau\right)} v_i + \frac{\exp\left(v_i^T x_i/\tau\right)}{\exp\left(v_q^T x_i/\tau\right)+\exp\left(v_i^T x_i/\tau\right)} x_i \tag{9}$$



(a) SeqNet (b) SeqXt

Figure 6: Prototype update scheme

3.2.3 Selective Memory Refreshment

The hard sample assistance strategy introduced in the previous section only deals with labeled samples, that is, persons with identity labeling This section aims to optimize the way the circular queue of storage is updated. During training, the circular queue is updated In the form of "First In First Out" (FIFO), as shown in Fig. 7a. However, the FIFO update strategy ignores the manifold structure of the vectors in the current data batch vs. the vectors stored in the circular queue. For example, vectors with high redundancy will be wrongly added to the circular queue, while vectors with high discrimination will be wrongly removed from the circular queue. Therefore, the vectors stored in the circular queue will become repetitive and trivial, which is not conducive to the learning of the network.

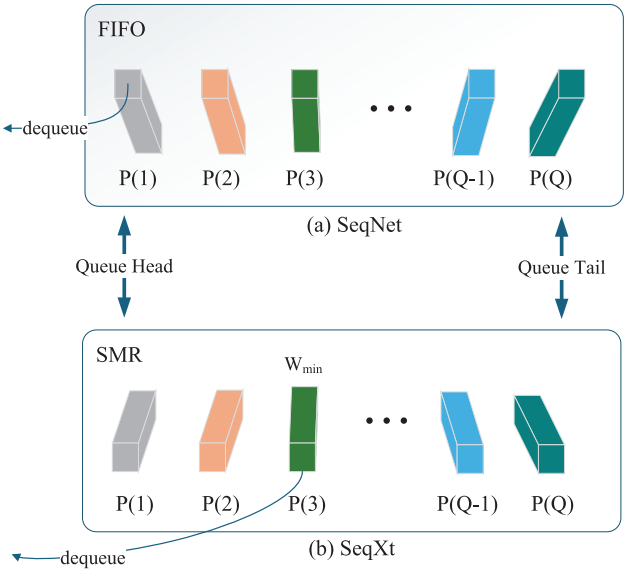


Figure 7: Difference between selective memory refresh and batch normalization

To alleviate this problem, the Selective Memory Refreshment (SMR) method proposed in this section, as shown in Fig. 7b, is proposed. It aims to rank each feature vector by assigning them an importance coefficient ω. Only if the ω of a vector in the current data batch is greater than the minimum ω of the vector stored in the circular queue, it will be added to the circular queue and replaced by the vector with the minimum ω. Specifically, ω should be designed to describe three properties of the feature vector:

1. Difficulty degree: For a person without identity labeling, if he has high similarity with the vector of persons with identity labeling in look-up table L, he is considered as a difficult sample that is difficult to distinguish. Difficult samples are more beneficial for model training than easy samples because they can provide a larger gradient magnitude to the training process.
2. Diversity: The feature vectors in the circular queue should be distinct from each other to act as non-trivial negative samples. Therefore, samples with high similarity to feature vectors in the existing cohort should be assigned lower importance.
3. Timeliness: As the model is constantly updated, the vectors in the memory matrix need to be updated accordingly. So the importance coefficient should decay with each iteration step of the model.

Based on the above three rules, this paper defines the importance coefficient of a person x with unmarked identity as follows.

$$\omega = \frac{\max\left(Lx\right)}{\max\left(Qx\right) + \varepsilon} \times k^{l} \tag{10}$$

where L represents the lookup table that stores the identified persons; k is a decay factor in the range $(0,1)$. l denotes the number of iterations; ε is a small constant used to ensure numerical stability. For any feature vector x, regardless of whether it comes from the person or the background, its importance coefficient ω is first calculated. The vector with the smallest importance coefficient is then found from the corresponding circular queue and compared with the candidate importance coefficient ω. If $\omega > \omega_{min}$, then x replaces the vector with the smallest importance coefficient in the circular queue. Otherwise, x is ignored and the circular queue is not updated at this point.

3.2.4 Training and Inference

The proposed network is supervised by the following five loss functions.

- L_{reg1}/L_{reg2}

$$L_{reg} = \frac{1}{N_p}\sum_{i=1}^{N_p} L_{loc}\left(r_i, \Delta_i\right) \tag{11}$$

- L_{cls1}

$$L_{cls1} = -\frac{1}{N}\sum_{i=1}^{N} c_i \log\left(p_i\right) \tag{12}$$

- L_{cls2}, L_{reid}

$$L_{cls2}, L_{reid} = L_{nae}(f) \tag{13}$$

Total loss function:

$$L = \lambda_1 L_{reg1} + \lambda_2 L_{cls1} + \lambda_3 L_{reg2} + \lambda_4 L_{cls2} + \lambda_5 L_{reid} \tag{14}$$

where $\lambda_1 = 10, \lambda_2 = \lambda_3 = \lambda_4 = \lambda_5 = 1$.

4 Dataset and Evaluation Protocol

CUHK-SYSU: The CUHK-SYSU [2] dataset is a hybrid dataset composed of urban scenes captured by mobile cameras and movie screenshots. It contains 18,184 uncropped images and 96,143 pedestrian bounding boxes. The training set includes 11,206 images and 55,272 pedestrians, while the test set consists of 2900 target pedestrians (query) and 6978 images for searching (gallery). The dataset filters out pedestrians based on pose and size; smaller pedestrians (height less than 50 pixels), heavily occluded ones, and those in complex poses (sitting, squatting, etc.) are not annotated. Pedestrians with variations in clothing and accessories are labeled as different categories, making person search in this dataset less challenging.

PRW: The PRW [4] dataset consists of video frames recorded by six fixed cameras installed at different locations on the Tsinghua University campus. It contains 11,816 frames and 43,110 bounding boxes, of which 34,304 bounding boxes belong to 932 identified pedestrians, with the rest marked as unidentified pedestrians. The training set includes 5704 images and 482 different identities. The test set comprises 2057 target pedestrians, who need to be searched among 6112 images. The varied camera viewpoints add complexity to this dataset. Although the number of pedestrian categories is smaller, each category contains a significant number of samples, making person search in this dataset more challenging.

UESTC-PS: The UESTC-PS (Person Search of University of Electronic Science and Technology of China) dataset is a private dataset created by the author's team of this paper. The data source consists of seven cameras placed sequentially in locations such as dormitory corridors, dormitory entrances, intersections, plazas in front of teaching buildings, building entrances, lobbies, and elevator exits within the Shenzhen Institute of Advanced Technology. The location of these cameras involves the road sections that students must pass through during class, and the capture scene is during the peak hours of students attending classes during the day. Due to the fact that this dataset is collected from real-life scenes and the location of the cameras is different, the dataset contains many difficulties in person search tasks, such as lighting, occlusion, posture, misalignment, viewing angles, and complex background environments. Thus, this dataset poses significant challenges for person search models, making it more difficult than the PRW dataset.

In addition, most publicly available pedestrian search datasets currently use multiple cameras facing the same scene to capture images of pedestrians in different poses. The background of the environment in which pedestrians are located changes relatively little, which is not in line with the changing camera angles in actual scenes. Therefore, we use pedestrian images with large background differences from multiple cameras. The collection and calibration process of the dataset fully meets the practical application scenarios, these characteristics are not present in publicly available datasets like CUHK-SYSU and PRW.

We selected 24,203 frames of relatively high quality from the original 164,074 frames to create our dataset. These video frames are divided into 588 video segments, containing a total of 533 different pedestrian IDs, with 127 repeated IDs appearing under different cameras. According to the characteristics of the person search task, training IDs can only appear in the training set, and test IDs can only appear in the test set. Therefore, we ultimately divided these 127 IDs into different training and test sets. The training set contains 295 video segments, 277 pedestrian IDs, and 74 repeated IDs. The test set contains 293 video segments, 256 pedestrian IDs, and 53 repeated IDs. The statistical comparison of the datasets is shown in Table 1.

Table 1: Comparison among different datasets

Data type	PRW	CUHK-SYSU	UESTC-PS
Frame	11,816	18,184	24,203
ID	932	8432	533
Annotated box	34,304	99,809	59,928
Box/ID	36.8	11.8	112.4
Train sequence	0	0	295
Test sequence	0	0	293
Train ID	482	5532	277
Test ID	450	2900	256
Camera	6	–	7
Type	Image	Image	Image + video
Monitoring scene	Single	Multi	Multi
Background change	Slight	Serious	Serious
Occlusion situation	Slight	Slight	Serious
Shooting angle	Smooth inspect	Smooth inspect	Overlook
Trajectories	No	No	Yes
Size	Big	Big	Big + small
Scenes	Outside	Outside	Outside + indoor

Evaluation Metric: For pedestrian detection performance, this paper uses Average Precision (AP) and Recall as metrics. A detected bounding box is considered a true positive if its Intersection over Union (IoU) with the annotated bounding box is higher than 0.5. For pedestrian search performance, this paper uses mean Average Precision (mAP) and top-K accuracy as metrics. mAP reflects the accuracy and match rate of searching for target individuals in the gallery. Top-K accuracy is widely used in pedestrian re-identification tasks. If at least one of the top-K predicted bounding boxes overlaps with the ground truth and has an IoU greater than or equal to a specified threshold, it is considered a correct match. In this paper, the IoU threshold is set to 0.5.

5 Experiments

We conducted experiments on the public pedestrian search dataset CUHK-SYSU and PRW, as well as the private dataset UESTC-PS, to demonstrate the effectiveness of the SeqXt network proposed in this paper, and compared it with state-of-the-art methods.

5.1 Implementation Details

The implementation of this paper is based on PyTorch. The backbone network is initialized with weights pre-trained on the ImageNet dataset. During training, the first convolutional layer of ConvNeXt is frozen and does not update. A prototype-guided normalization layer is added after the fully connected layer that generates the pedestrian feature vectors on top of the region convolutional network. The dimension (d) of the generated feature vectors is set to 256.

The softmax temperature and the decay factor k for the importance coefficient in the dynamic online instance matching (DOIM) loss function are set to 1/30 and 0.99, respectively. The size of the memory matrix, L and Q, is set differently according to the dataset. For CUHK-SYSU, they are set to 5532 and 5000, respectively; for PRW, L is set to 482, and Q is reduced to 500 to balance the number of classes; for the UESTC-PS dataset, N and M are set to 534 and 500, respectively.

The model is trained end-to-end on a single NVIDIA Tesla P40 GPU. The batchsize is set to 5, and each input image is resized so that its shorter side is at least 900 pixels and its longer side is at most 1500 pixels. When necessary, cropping and zero-padding are used to fit images of different resolutions into a batch. The learning rate is gradually increased during the first epoch, starting from 0.00015 and increasing to 0.003, then it is maintained until it decays to 0.1 of its original size at the 16th epoch, and training ends after the 22nd epoch.

5.2 Experimental Results and Analysis

This article conducted detailed experiments on public pedestrian search datasets CUHK-SYSU, PRW, and private datasets UESTC-PS to verify the effectiveness of the proposed improvement strategy and compare its performance with state-of-the-art algorithms.

5.2.1 Comparison of Different Backbone

By observing the data in Table 2, we can see that the SeqXt model demonstrates significant advantages over the Baseline model across all performance metrics. Specifically, the SeqXt model surpasses the Baseline model in key metrics such as Recall, AP, mAP, and top-1, top-5, and top-10 accuracies. The SeqXt model uses ConvNeXt as the backbone network, and compared to the Resnet50 used by the Baseline model, it shows superior performance. This fully proves the effectiveness of ConvNeXt in feature extraction and image classification tasks.

Table 2: Comparison of experimental results of different backbone networks on the PRW dataset

Model	Backbone	Recall (%)	AP (%)	mAP (%)	Top-1 (%)	Top-5 (%)	Top-10 (%)
SeqNet	Resnet50	96.6	94.1	46.1	83.1	92.1	94.5
SeqXt	ConvNeXt	96.5	94.1	53.3	86.5	93.7	95.3

Furthermore, the SeqXt model achieves particularly outstanding results in pedestrian re-identification experiments on the PRW dataset, with a mean average precision (mAP) reaching 53.3%, compared to 46.1% for the baseline linear model, achieving a significant performance improvement. The advantages demonstrated by ConvNeXt as the backbone network in pedestrian search and identification tasks provide new research directions for facing more complex pedestrian search and identification tasks in the future, showing broad application prospects.

5.2.2 Effects of Various Components in DOIM

We selected SeqXt as the benchmark model, which achieved impressive scores of 53.3% in mean average precision (mAP) and 86.5% in top-1 accuracy, laying a solid foundation for subsequent experiments. To further explore the effectiveness of various components in the dynamic online instance matching (DOIM) loss proposed in this chapter, we conducted ablation experiments on the PRW dataset. First, we integrated the hard example auxiliary strategy (HEA) into the benchmark model. Compared to the baseline model, the introduction of the HEA module resulted in an increase of 0.5% in mAP and 0.2% in top-1 accuracy, strongly validating the positive effect of the hard example auxiliary strategy in enhancing model performance. Next, we replaced the original FIFO mechanism in the online instance matching circular queue with the selective memory refresh strategy. According to the experimental data in Table 3, this improvement led to an increase of 1.9% in mAP and 0.1% in top-1 accuracy, revealing the effectiveness of the selective memory refresh strategy in optimizing pedestrian matching. Finally, we organically integrated the various components proposed in this chapter to construct a complete DOIM loss. The experimental results showed that at this time, mAP and top-1 accuracy were improved by 2.0% and 0.7%, respectively. This outstanding performance fully validates the positive role of the DOIM loss proposed in this paper in pedestrian search tasks.

Table 3: Ablation experiments on the PRW dataset

Method	Recall (%)	AP (%)	mAP (%)	Top-1 (%)	Top-5 (%)	Top-10 (%)
SeqXt	96.5	94.1	53.3	86.5	93.7	95.3
+HEA	96.6	94.0	53.7	86.7	93.6	94.8
+SMR	96.6	94.2	54.2	86.4	93.8	95.0
Ours(DOIM)	96.7	94.4	55.3	87.2	94.0	95.1

5.2.3 Effects of Different Combination Strategies

We designed ablation experiments on the PRW dataset to investigate the effects of different normalization methods and loss function combinations on pedestrian search performance. The experimental results are shown in Table 4.

Table 4: Ablation experiments of different strategy combinations on the PRW dataset

BN	PN	L_{OIM}	L_{DOIM}	Person search recall (%)	Person search AP (%)	Re-ID mAP (%)	Re-ID top-1 (%)
✗	✗	✓	✗	95.8	92.5	44.9	84.4
✓	✗	✓	✗	96.8	94.2	53.3	86.5
✗	✓	✓	✗	97.1	94.1	54.8	87.6

(Continued)

Table 4 (continued)

BN	PN	L_{OIM}	L_{DOIM}	Person search recall (%)	Person search AP (%)	Re-ID mAP (%)	Re-ID top-1 (%)
✓	✗	✗	✓	96.8	94.0	55.3	87.2
✗	✓	✗	✓	97.3	94.5	56.5	88.3

Firstly, When no normalization method is used and only the OIM loss function is employed, the model's mAP and top-1 accuracy are 44.9% and 84.4%, respectively. This relatively low result indicates that normalization plays an important role in pedestrian search tasks. Next, we introduced the batch normalization method and observed a significant improvement in model performance. Specifically, mAP and top-1 accuracy increased to 53.3% and 86.5%, respectively, with increases of 8.4% and 2.1%. This result validates the effectiveness of batch normalization in improving the performance of pedestrian search models.

Then, we tried using the prototype-guided normalization method proposed in this paper to replace batch normalization. Experimental results show that this replacement brought additional performance improvements, with mAP and top-1 accuracy reaching 54.8% and 87.6%, respectively. This indicates that the prototype normalization method has advantages in calibrating feature distribution and can further improve the accuracy of pedestrian search. In addition, we tested the effect of the dynamic online instance matching (DOIM) loss proposed in this paper. Compared to using only the OIM loss function, the introduction of the DOIM loss increased mAP and top-1 accuracy to 55.3% and 87.2%, respectively. This result proves the effectiveness of the hard example auxiliary update strategy and the selective memory refresh strategy in optimizing the online instance matching function.

Finally, we combined the prototype normalization method and the DOIM loss and obtained the best performance. Specifically, the model's mAP and top-1 accuracy reached 56.5% and 88.3%, respectively, representing increases of 11.6% and 3.9% compared to using only the online instance matching function. These results fully demonstrate the superior performance of the combination of the normalization method and the loss function proposed in this paper in pedestrian search tasks.

5.2.4 Performance on UESTC-PS

In this study, we further conducted experimental evaluations of the proposed network model on the private dataset UESTC-PS. Given that the UESTC-PS dataset is derived from real-life scenarios and involves multiple cameras from different locations, it contains various challenging factors such as lighting changes, occlusion, pose differences, misalignment issues, diverse observation angles, and complex background environments, making it particularly representative in pedestrian search tasks. Therefore, the evaluation of pedestrian search models using the UESTC-PS dataset is significantly more challenging than using the PRW dataset, providing a more stringent test criterion for the performance of the model.

Based on the detailed experimental data in Table 5, we can clearly observe that the pedestrian search model proposed in this chapter exhibits significant performance advantages, not only surpassing current advanced network models, but also achieving 71.1% and 81.1% accuracy in mean Average Precision (mAP) and top-1 accuracy, respectively. It is worth mentioning that compared with the baseline network, this model has improved in all indicators, especially in terms of overall average accuracy and top-1 accuracy, achieving growth of 3.9% and 4.8%, respectively. This result fully demonstrates the effectiveness of the proposed model in extracting discriminative pedestrian representations, as well as its strong robustness. Fig. 8 illustrates the

trends in mAP for SeqNet and our model, from which the superiority of our model is evident. The excellent performance on the challenging private dataset UESTC-PS further validates the effectiveness of the proposed improvement strategy in practical applications.

Table 5: Experimental results of different algorithms on the UESTC-PS dataset

Method	Source	Recall (%)	AP (%)	mAP (%)	Top-1 (%)	Top-5 (%)	Top-10 (%)
COAT	CVPR 2022	96.4	94.4	68.7	77.6	96.0	98.8
PSTR	CVPR 2022	96.2	94.4	68.4	76.9	95.9	98.5
SeqNet	AAAI 2021	96.5	94.8	67.2	76.3	95.8	98.50
Ours	–	97.0	95.4	71.1	81.1	96.5	98.50

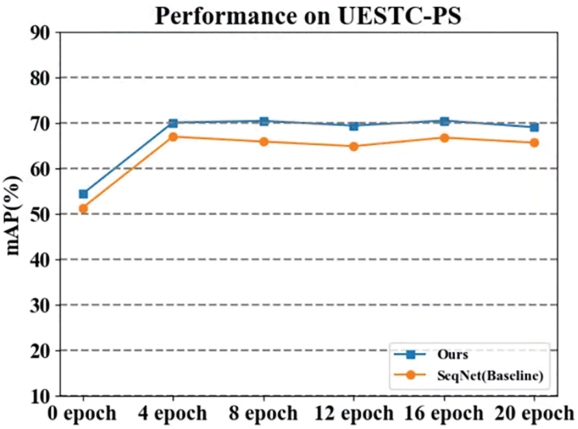


Figure 8: Variation curve of mAP with epoch

5.2.5 Efficiency Comparison

Based on experiments with the PRW dataset and a Quadro RTX 8000 GPU, we compared the efficiency of SeqNet and SeqXt in Table 6. Since SeqXt leverages ConvNext to extract pedestrian features, its network structure is more complex than the ResNet50 used in SeqNet, leading to significantly higher parameter counts, FLOPs, and lower FPS. Although efficiency is slightly reduced, the improved pedestrian feature accuracy from ConvNext-Base enables SeqXt to achieve a much higher mAP than SeqNet.

Table 6: Efficiency comparison

Method	Param (M)	FLOPs (G)	FPS	mAP (%)
SeqNet	48.42	555.05	16.7	46.7
SeqXt(ours)	122.68	751.55	11.4	56.55

5.3 Visualization Analysis

To validate the search performance of SeqXt across diverse scenarios, we use the UESTC-PS dataset for visualization analysis, as shown in Fig. 9. With the exception of the Fig. 9c, all detection results marked with green bounding boxes indicate a similarity exceeding 0.5.

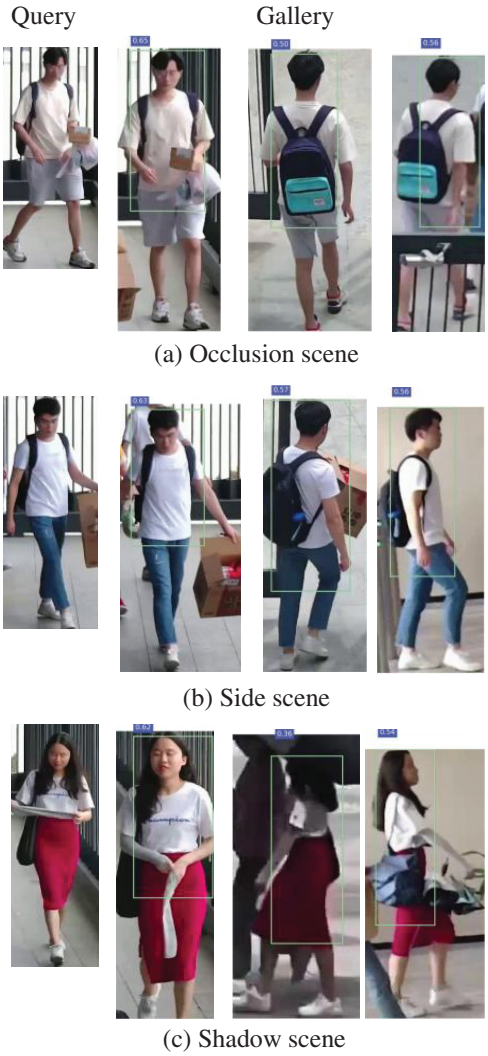


Figure 9: Visualization results of SeqXt

The UESTC-PS dataset is collected from real-world multi-camera surveillance scenarios, encompassing challenging factors such as illumination variations, occlusions, posture differences, misalignment issues, multi-viewpoint angles, and complex backgrounds. These attributes render it highly representative for person search tasks. Leveraging the powerful feature extraction capability of the ConvNext backbone network, SeqXt effectively captures critical target features. Visualization results demonstrate that the model maintains robust performance under most challenging conditions, verifying its overall effectiveness.

As shown in Fig. 9a, targeting a student from the Graduate School, the model achieved precise identification despite significant apparel changes and fence occlusion. As shown in Fig. 9b, targeting another student, the model completed retrieval with high confidence even when only a side-view profile was exposed. As shown in Fig. 9c, targeting a faculty member, severe head occlusion combined with shadow interference led to a significant drop in recognition accuracy, indicating that the model's adaptability to extreme lighting conditions requires further optimization.

5.4 Experimental Results and Analysis

To highlight the superiority of the model in this chapter, we compared the performance of our method with other algorithms on the CUHK-SYSU dataset. The experimental data is detailed in Table 7, where the size of the search library is uniformly set to 100. This chapter's algorithm incorporates the previously proposed improvement strategy based on SeqNet, and the results show that the algorithm achieved excellent results of 96.1% and 96.5% in mAP and top-1 metrics, respectively. Compared with the baseline model, the performance has been significantly improved, with mAP and top-1 improving by 2.3% and 1.9%, respectively, surpassing many advanced algorithms. This experimental result fully proves that the performance improvement of the algorithm in this chapter is significant, and the series of improvement measures taken are scientific and effective.

Table 7: Comparison with state-of-the-art methods

Method	CUHK-SYSU mAP (%)	CUHK-SYSU top-1 (%)	PRW mAP (%)	PRW top-1 (%)
Two-step CLSA [5]	87.2	88.5	38.7	65
End-to-end ASTD [24]	95.8	96.2	55.7	90.2
Two-step IGPN [25]	90.3	91.4	42.9	70.2
Two-step TCTS [26]	93.9	95.1	46.8	87.5
End-to-end OIM [2]	75.5	78.7	21.3	49.9
End-to-end NAE [10]	91.5	92.4	43.3	80.9
End-to-end SeqNet [11]	93.8	94.6	46.7	83.4
End-to-end MQPS [17]	94.23	94.07	52.20	87.86
End-to-end DTNN [19]	94.9	95.3	51.6	87.6
End-to-end PSTR [20]	93.5	95.0	49.5	87.8
End-to-end COAT [21]	94.2	94.7	53.3	87.4
End-to-end AlignPS [27]	93.1	96.4	45.9	81.9
Ours	96.1	96.5	56.5	88.3

We further tested the performance of different algorithms on the PRW dataset and summarized the results in Table 7. After comparative analysis, the algorithm proposed in this chapter demonstrates excellent performance in both mAP and top-1 indicators, reaching the industry's advanced level and successfully surpassing most algorithms. Compared with the benchmark model, the performance of the algorithm in this chapter has been significantly improved, with mAP and top-1 improving by 9.8% and 4.9%, respectively, ultimately reaching high levels of 56.5% and 88.3%. The achievement of this experimental result fully proves the rationality and effectiveness of the improvement strategy adopted in this chapter, providing a more accurate and reliable method for pedestrian search tasks.

Empirical evidence reveals that the integration of ConvNeXt-Base as the feature extraction backbone into the SeqNet architecture enables meticulous extraction of pedestrian feature embeddings, effectively mitigating the inherent limitation of ResNet50 in resolving fine-grained details under complex scene dynamics. The Prototype-Guided Normalization mechanism augments the discriminative power of pedestrian features by rectifying the class-wise representation bias, thereby addressing the inter-class discrimination asymmetry prevalent in state-of-the-art approaches. Within the DOIM framework, the HEA module strategically amplifies the weighting schema for challenging instances, while the SMR module refines the circular queue

dequeue protocol, fundamentally revamping the hard sample and circular queue updating mechanisms of traditional OIM.

6 Conclusion

This article uses advanced ConvNeXt-Base as a feature extraction network to extract more discriminative features, thereby capturing subtle differences in pedestrian recognition tasks and improving recognition accuracy. At the same time, in order to balance the class distribution in the dataset and enhance the intra class compactness and inter class separability of pedestrian IDs, the prototype guided normalization (PN) method is introduced, which calibrates the feature distribution through prototype vectors to ensure that the features do not excessively lean towards frequently occurring IDs. In addition, in order to further enhance the model's generalization ability and recognition ability for difficult samples, this paper proposes a novel loss function—Dynamic Online Instance Matching(DOIM) Loss. DOIM adaptively updates the lookup table (LUT) and cyclic queue (CQ) to capture difficult samples in real-time during the training process, thereby promoting the enhancement of inter class feature differences. This dynamic update mechanism enables the model to pay more attention to difficult to distinguish samples, improving the model's recognition ability for difficult samples. The results on public datasets CUHK-SYSU, PRW, and private datasets UESTC-PS show that our method has achieved better experimental results than the baseline model. Beyond pedestrian search, the proposed model demonstrates substantial potential for cross-domain adaptive tracking scenarios, including suspected vehicle search, traffic violation systems, and other intelligent transportation applications.

Acknowledgement: We would like to acknowledge the Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China (UESTC) for providing camera data collection support. We also extend our gratitude to the Public Safety Technology Research Institute of UESTC for their GPU equipment support during this research.

Funding Statement: This work was supported by the major science and technology special projects of Xinjiang (No. 2024B03041) and the scientific and technological projects of Kashgar (No. KS2024024).

Author Contributions: The authors confirm their contribution to the paper as follows: study conception and design: Xiuchuan Cheng; data collection: Meiling Wu; analysis and interpretation of results: Xiuchuan Cheng and Xu Feng; draft manuscript preparation: Xiuchuan Cheng, Meiling Wu, Xu Feng, Zhiguo Wang, Guisong Liu and Ye Li. All authors reviewed the results and approved the final version of the manuscript.

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

Ethics Approval: The private dataset UESTC-PS involved in this study was collected with the informed consent of the participants, and thus entails no relevant ethical risks.

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

References

1. Xu Y, Ma B, Rui H, Lin L. Person search in a scene by jointly modeling people commonness and person uniqueness. In: MM '14: Proceedings of the 22nd ACM International Conference on Multimedia; 2014 Nov 3–7; Orlando, FL, USA. p. 937–40. doi:10.1145/2647868.2654965.
2. Xiao T, Li S, Wang B, Lin L, Wang X. Joint detection and identification feature learning for person search. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 3415–24.
3. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2017;39(6):1137–49. doi:10.1109/tpami.2016.2577031.

4. Zheng L, Zhang H, Sun S, Chandraker M, Yang Y, Tian Q. Person re-identification in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 1367–76.
5. Lan X, Zhu X, Gong S. Person search by multi-scale matching. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018 Sep 8–14; Munich, Germany. p. 536–52.
6. Han C, Ye J, Zhong Y, Tan X, Zhang C, Gao C, et al. Re-id driven localization refinement for person search. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 9814–23.
7. Yan Y, Zhang Q, Ni B, Zhang W, Xu M, Yang X. Learning context graph for person search. arXiv:1904.01830. 2019.
8. Munjal B, Amin S, Tombari F, Galasso F. Query-guided end-to-end person search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019 Jun 15–20; Long Beach, CA, USA. p. 811–20.
9. Zhang Y, Yang Y, Kang W, Zhen J. Multi-scale occlusion suppression network for occluded person re-identification. Pattern Recognit Lett. 2024;185(C):66–72. doi:10.1016/j.patrec.2024.07.009.
10. Chen D, Zhang S, Yang J, Schiele B. Norm-aware embedding for efficient person search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 12615–24. doi:10.1109/CVPR42600.2020.01263.
11. Li Z, Miao D. Sequential end-to-end network for efficient person search. Proc AAAI Conf Artif Intell. 2021;35(3): 2011–9. doi:10.1609/aaai.v35i3.16297.
12. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
13. Li Y, He J, Zhang T, Liu X, Zhang Y, Wu F. Diverse part discovery: occluded person re-identification with part-aware transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 20–25; Nashville, TN, USA. p. 2898–907.
14. Wang H, Shen J, Liu Y, Gao Y, Gavves E. Nformer: robust person re-identification with neighbor transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. p. 7297–307.
15. Zhang G, Zhang P, Qi J, Lu H. Hat: hierarchical aggregation transformers for person re-identification. In: MM '21: Proceedings of the 29th ACM International Conference on Multimedia; 2021 Oct 20–24; Online. p. 516–25. doi:10.1145/3474085.3475202.
16. Chen L, Xu J. Sequential transformer for end-to-end person search. In: International Conference on Neural Information Processing. Singapore: Springer Nature Singapore; 2023. p. 226–38.
17. Chen Y, Li Z, Song A. Multi-query person search with transformers. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Singapore: Springer Nature Singapore; 2024. p. 116–28. doi:10.1007/978-981-97-2238-9_9.
18. Lv N, Xiang X, Wang X, Qiao Y, El Saddik A. Learning feature contexts by transformer and CNN hybrid deep network for weakly supervised person search. Comput Vis Image Underst. 2024;239(5):103906. doi:10.1016/j.cviu.2023.103906.
19. Feng C, Han D, Chen C. DTHN: dual-transformer head end-to-end person search network. Comput Mater Contin. 2023;77(1):245–61. doi:10.32604/cmc.2023.042765.
20. Cao J, Pang Y, Anwer RM, Cholakkal H, Xie J, Shah M, et al. PSTR: end-to-end one-step person search with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. p. 9458–67.
21. Yu R, Du D, LaLonde R, Davila D, Funk C, Hoogs A, et al. Cascade transformers for end-to-end person search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. p. 7267–76.
22. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. arXiv:2005.12872. 2020.
23. Cai Z, Vasconcelos N. Cascade R-CNN: delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 6154–62.

24. Zhang Q, Miao D, Zhang Q, Wang C, Li Y, Zhang H, et al. Learning adaptive shift and task decoupling for discriminative one-step person search. Knowl Based Syst. 2024;304(4):112483. doi:10.1016/j.knosys.2024.112483.
25. Dong W, Zhang Z, Song C, Tan T. Instance guided proposal network for person search. In: Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 2582–91. doi:10.1109/CVPR42600.2020.00266.
26. Wang C, Ma B, Chang H, Shan S, Chen X. TCTS: a task-consistent two-stage framework for person search. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 11949–58. doi:10.1109/CVPR42600.2020.01197.
27. Yan Y, Li J, Qin J, Bai S, Liao S, Liu L, et al. Anchor-free person search. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 7686–95.