Context-aware Dynamic Contrastive Learning Network and E-Bike Rider Benchmark for Person Search

Hongchao Li[1], Chengcheng Li[1], Xixi Wang[2*], YongLong Luo[1]
[1]School of Computer and Information Science, Anhui Normal University
Wuhu, Anhui Province 241002, China
[2]School of Computer Science and Technology, Anhui University
Hefei, Anhui Province 230601, China
lhc950304@foxmail.com, licc@ahnu.edu.cn, sissiw0409@foxmail.com, ylluo@ustc.edu.cn

Abstract

Person search is a challenging computer vision task that aims to simultaneously detect and re-identify individuals from uncropped gallery images. However, most existing approaches are limited by restricted receptive fields, leading to distorted local feature representations under occlusions or complex poses. Additionally, scale variations hinder model generalization in real-world scenarios. To address these limitations, we introduce a novel E-Bike Rider Search (EBRS) dataset, which comprises 27,501 images capturing 963 distinct IDs across 8 camera views at a large urban intersection in a Chinese city. Furthermore, we propose a Context-aware Dynamic Contrastive Learning (CDCL) framework that dynamically adjusts convolutional weights and performs hard sample mining based on contextual cues, thereby improving discriminative capability for both local details and global features. Extensive experiments show our method achieves state-of-the-art performance on CUHK-SYSU and PRW benchmarks, with competitive results on the challenging EBRS dataset, demonstrating its effectiveness.

Code & Datasets — https://github.com/licc3996/CDCL

Introduction

Person search, which involves simultaneously localizing and identifying target individuals in cluttered scene images, has emerged as a critical computer vision task. This technology has garnered substantial research interest due to its broad applications in intelligent surveillance systems, social media analytics, and public security domains. The evolution of person search methodologies has progressed through two primary paradigms: (1) Traditional two-stage approaches (Zheng et al. 2017; Wang et al. 2020) that separately perform person detection and re-identification, and (2) Contemporary end-to-end frameworks (Li and Miao 2021; Yan et al. 2021; Yu et al. 2022) that integrate these components into a unified network architecture, achieving superior computational efficiency. Despite these advancements, the field continues to confront significant challenges, particularly in handling: occlusions, pose variations and scale disparities.

Existing mainstream person search datasets (e.g., PRW (Zheng et al. 2017) and CUHK-SYSU (Xiao et al. 2017)) have laid the foundation for the field's development. However, as shown in Figure 1, their constructed scenarios exhibit some limitations: 1) Monotonous Scenarios: These datasets primarily consist of static environments (e.g., campuses, street photography, movie clips), lacking dynamic traffic interactions. 2) Insufficient Occlusion Challenges: The existing occlusions fail to address multi-level occlusions (vehicle-rider, inter-pedestrian, and partial occlusions) prevalent in dense urban scenes. 3) Limited scale diversity: Person images are mainly limited to medium-to-close ranges, with rare examples of extremely small-scale targets or non-rigid deformations caused by riding postures. To address these limitations, we propose EBRS (E-Bike Rider Search), the first large-scale benchmark for person search in urban e-bike scenarios. The dataset contains 27,501 images (963 identities across 8 cameras), captured during 16:00–18:00 at a busy intersection in a Chinese city. The dataset advances the field by intensifying three fundamental challenges: multi-source occlusions, extreme scale variations, and complex pose deformations. By capturing authentic urban dynamics, EBRS propels person search research toward more realistic and challenging environments, serving as an essential benchmark for developing robust algorithms in real-world applications.


(a) (b)

Figure 1: (a) Existing datasets: static scenes (campuses, streets, movie clips). (b) EBRS dataset: complex urban e-bike scenarios.

Real-world person search faces three representation-level challenges, including occlusions, pose variations, and scale disparities. These combined challenges create substantial difficulties for traditional Convolutional Neural Networks (CNNs), which rely on fixed-size convolution kernels and

*Corresponding author.

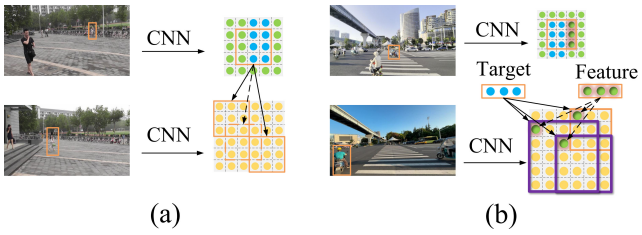

Figure 2: (a) Fixed kernels: limited for occlusion/pose/scale variations. (b) Context-aware dynamic kernels: learnable spatial weights conditioned on scene context.

local receptive fields, as shown in Figure 2 (a). Regarding occlusions, fixed receptive fields lack global contextual awareness, failing to reconstruct heavily obscured body regions. For pose variations, rigid convolution kernels cannot dynamically adjust to articulated deformations, resulting in severe part misalignment. Concerning scale disparities, small targets suffer detail loss from oversized kernels while large targets lose structural relationships due to constrained receptive fields. To this end, we propose a novel Context Mixed Convolution (CMC) module. First, the CMC module introduces a dynamic feature learning mechanism that adaptively fuses multi-scale contextual information. Second, it hierarchically aggregates channel-wise and spatial-wise dependencies through dual-path calibration, followed by cross-scale feature integration while preserving spatial-channel consistency. Third, CMC generates context-aware dynamic kernels that automatically adjust receptive fields based on contextual cues, enabling simultaneous local detail preservation and global structure modeling, as shown in Figure 2 (b). Finally, the cascaded refinement design progressively enhances feature discriminability through two-stage contextual dynamic aggregation, achieving robust representation learning under complex real-world conditions.

In addition, discrimination-level challenges arise from intra-class variance and inter-class similarity. The feature distribution of the same person exhibits large variations due to occlusion, pose, and scale changes, while features of different persons may become indistinguishable due to similar clothing or carried items. In conventional contrastive learning frameworks, which rely on binary classification (positive/negative samples), the inclusion of hard negatives in regular negative samples introduces training bias and fails to resolve these issues effectively. The key limitations lie in handling hard negatives and dynamically adjusting sample weights. To overcome the dual challenges of significant intra-class variations and high inter-class similarity in person search, we propose a Dynamic Contrastive Learning (DCL) module. The DCL module introduces a triplet-category learning paradigm (i.e., positive, easy-negative, hard-negative) that restructures feature label relationships. Through dynamic loss weighting, it simultaneously enhances discriminative learning of hard negatives while maintaining intra-class compactness of positives, improving network robustness.

Our contribution can be summarized as follows:

- To our knowledge, we are the first to construct the E-Bike Rider Search (EBRS) dataset targeting complex urban scenes, which improves the applicability of person search models in real-world scenarios.
- We propose a Context-aware Dynamic Contrastive Learning (CDCL) framework that synergistically combines context-aware dynamic convolution with contrastive learning, effectively handling occlusion/pose/scale variations while enhancing feature discrimination.
- Comprehensive experiments on PRW, CUHK-SYSU and EBRS datasets validate the superiority of our model, especially in challenging urban surveillance conditions.

Related Work

Person search methods are broadly categorized into two paradigms: two-stage and end-to-end approaches.

Two-stage methods. Early works (Xu et al. 2014; Chen et al. 2018) treat detection (Ren et al. 2016) and re-identification (Luo et al. 2019) as separate tasks. For instance, (Wang et al. 2020) propose the TCTS method to address inconsistencies between detection and ReID. (Lan, Zhu, and Gong 2018) introduce a multi-scale feature pyramid for person detection and ReID, while (Zheng et al. 2017) adopt a two-stage fine-tuning strategy. (Huang et al. 2023) develop the CFAR method, leveraging pose and parsing priors to mitigate feature misalignment. Despite their contributions, these methods suffer from pipeline fragmentation and computational redundancy.

End-to-end methods. (Xiao et al. 2017) first propose joint learning of detection and ReID using a unified backbone. Subsequently, many end-to-end methods emerge (Munjal et al. 2019; Chen et al. 2020a,b; Dong et al. 2020a; Han, Ko, and Sim 2021). Nevertheless, the inherent conflict between detection's need for generalization and ReID's requirement for discrimination persists as a key challenge (Xu et al. 2014). Recent solutions include: staged optimization (Li and Miao 2021), task decoupling (Han et al. 2023), and joint pretraining (Tian et al. 2024). (Wang et al. 2024) incorporate intra-image contrastive learning, whereas (Yan et al. 2021) leverage deformable convolutions for spatial alignment. (Yu et al. 2022) propose a Closed-loop Attention Transformer that enhances higher-level alignment. (Fiaz et al. 2023) present a Scale Adjustment Transformer that effectively integrates multi-granularity features. (Jiang et al. 2024) propose SEAS, which suppresses background and foreground noise through bilateral modulations to maintain representation consistency. (Kim, Lee, and Sohn 2025) introduce PAD, boosting re-identification discriminability via prototype-guided attention distillation at the cost of higher computational overhead. Different from end-to-end methods, our work introduces both the EBRS dataset and a CDCL model specifically designed for e-bike rider search with occlusions and pose variations.

E-Bike Rider Benchmark

To enhance scenario diversity in person search datasets, we propose the first dedicated benchmark dataset for E-Bike riders, termed the E-Bike Rider Search (EBRS) Dataset.

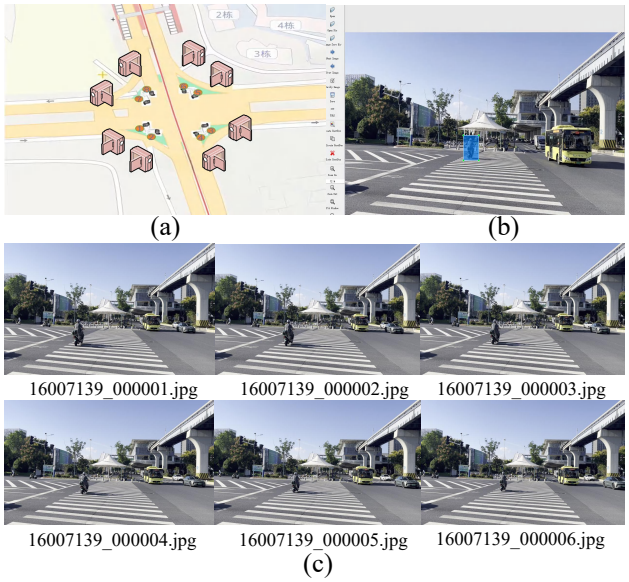

Figure 3: Data collection and annotation: (a) Camera setup with eight synchronized locations; (b) Manual bounding box annotation; (c) Frame sampling at 6 fps.

This section details the data acquisition process, followed by an in-depth discussion of the dataset construction methodology and the challenges encountered during its development.

Data Collection. The dataset collection occurs during 16:00–18:00 in September at a strategically selected intersection in a representative medium-sized Chinese city. To comprehensively capture the visual attributes of e-bike riders, we deploy eight synchronized cameras at distinct viewpoints (as shown in Figure 3 (a)), ensuring complete coverage of passing riders. For each camera pair, critical parameters including focal length, aperture, and exposure time are rigorously calibrated to ensure maximal consistency in lighting conditions and perspective across all image captures. All cameras capture 1080p video at 30 fps.

Data Annotation. As shown in Figure 3 (b, c), we first extract frames from the videos captured by each camera. To ensure data diversity while avoiding redundancy, we sample each video at 6 fps to construct the raw dataset. For every e-bike appearing in the raw frames, we manually annotate precise bounding boxes to localize their positions. By leveraging multiple cues including timestamps, camera locations, vehicle positions, and surrounding context, we associate frames containing the same e-bike rider to ensure accurate identity assignment. Following annotation, we perform manual data cleaning to verify and correct potential labeling errors or incomplete information, thereby ensuring dataset accuracy and reliability. Finally, we organize the cleaned data into the final dataset based on camera ID, time segment ID, image ID, and rider ID.

Data Description. Through manual annotation of 27,501 frames, we obtain 35,136 rider bounding boxes while successfully identifying and labeling 963 distinct IDs. For effective model training and evaluation, the dataset is partitioned into training and test subsets. The training set comprises 19,385 images (24,742 bounding boxes) representing 626 unique rider IDs, while the test set contains 8,116 images (10,394 gallery boxes, 5,387 query boxes) with 337 different IDs. Consistent with PRW (Zheng et al. 2017), we use all non-query test images as the gallery for evaluation.

Methodology

We propose a Context-aware Dynamic Contrastive Learning (CDCL) framework for robust person search in complex scenarios, as illustrated in Figure 4. Building on standard backbones like ResNet50 (He et al. 2016), our framework introduces two key modules: a Context Mixed Convolution (CMC) module that employs context-aware dynamic kernels to adaptively learn target features, and a Dynamic Contrastive Learning (DCL) module that improves feature discrimination through optimized hard negative weighting.

Context Mixed Convolution (CMC)

Person search suffers from three core challenges: occlusions, pose variations, and scale disparities, which render local features insufficient for reliable identification. Traditional CNNs exacerbate this by progressively losing spatial details through pooling operations. Our solution introduces a Context Mixed Convolution (CMC) module, as shown in Figure 4 (a), that strategically acquires and leverages multi-scale contextual information to overcome these limitations.

Multi-scale context extraction. Given an input image, layers 1-3 of the bockbone network serve as the feature extractor, producing the multi-scale feature maps $\mathbf{F}_k$ ($k \in \{1, 2, 3\}$). Then, global average pooling (GAP) is applied to obtain the corresponding channel descriptors. To effectively capture both channel-wise and spatial-wise contextual dependencies, we process these features through:

$$\begin{aligned} c_k &= \sigma\left(\text{Linear}\left(\text{GELU}\left(\text{Linear}(\text{GAP}(\mathbf{F}_k))\right)\right)\right), \\ s_k &= \sigma\left(\text{Conv}_{3\times3}(\mathbf{F}_k)\right), \end{aligned} \tag{1}$$

where Linear and $\text{Conv}_{3\times3}$ denote linear and convolutional layers respectively, with GELU and σ representing the Gaussian error linear unit and sigmoid activation functions. The calibrated features are obtained via: $\bar{\mathbf{F}}_k = \mathbf{F}_k \odot c_k \odot s_k$, where $\odot$ denotes element-wise multiplication of channel (c_k) and spatial (s_k) attention weights with original features.

To effectively learn multi-scale contextual information, we integrate features from different stages through the following operation: $\bar{\mathbf{F}} = \bar{\mathbf{F}}_3 + \text{Up}(\text{Conv}_{1\times1}(\bar{\mathbf{F}}_1)) + \text{Up}(\text{Conv}_{1\times1}(\bar{\mathbf{F}}_2))$, where $\text{Conv}_{1\times1}$ unifies channel dimensions. $\text{Up}(\cdot)$ aligns spatial resolutions via bilinear interpolation. The resulting fused representation $\bar{\mathbf{F}}$ comprehensively captures hierarchical features from all scales, maintaining both channel consistency and spatial alignment. The feature map $\bar{\mathbf{F}}$ then undergoes multi-scale context refinement to obtain the final contextual representation:

$$\begin{aligned} \mathbf{F}_{\text{up}} &= \text{Up}\left(\text{GELU}\left(\text{BN}\left(\text{Conv}_{3\times3}(\bar{\mathbf{F}})\right)\right)\right), \\ \mathbf{F}_{\text{ctx}} &= \text{BN}\left(\text{Conv}_{1\times1}(\mathbf{F}_{\text{up}})\right), \end{aligned} \tag{2}$$

where the strided 3×3 convolution enables cross-scale context interaction through downsampling, the 1×1 convolution

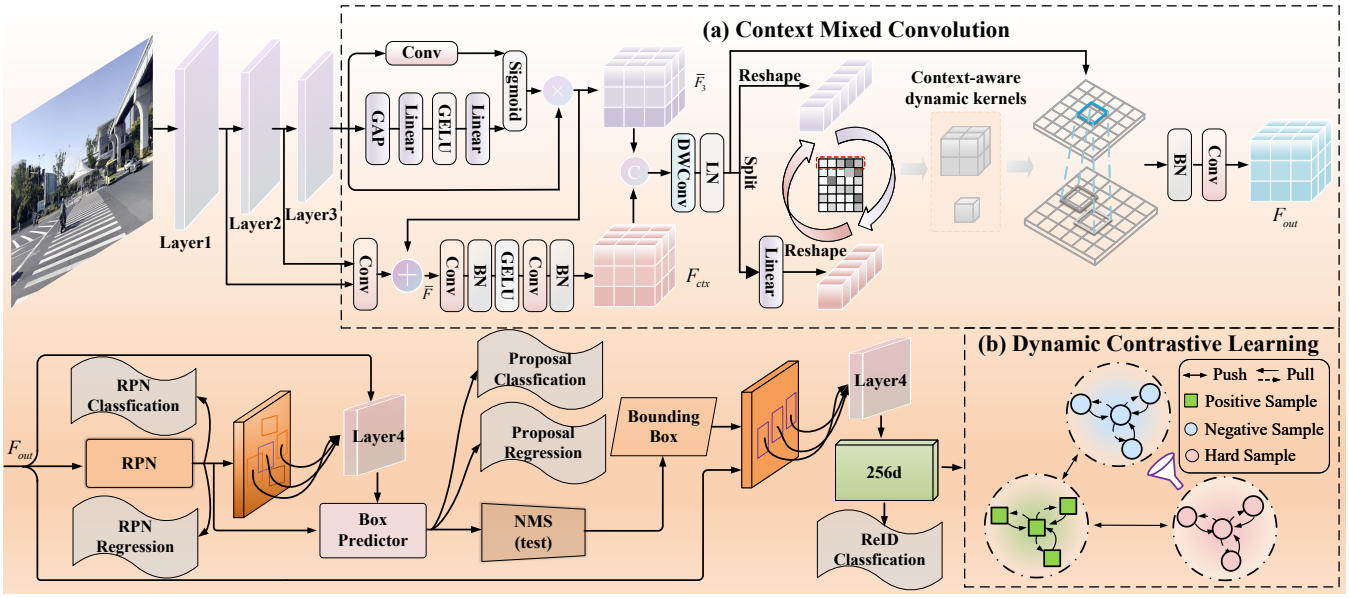

Figure 4: Architecture overview of our Context-aware Dynamic Contrastive Learning (CDCL) framework.

and upsampling operator ensure spatial and channel compatibility with the backbone features.

Context-aware Dynamic Aggregation. To fully leverage learned multi-scale contextual information, we design a Context-aware Dynamic Aggregation (CDA) operation that systematically fuses $\bar{\mathbf{F}}_3$ and $\mathbf{F}_{\text{ctx}}$:

$$\begin{aligned} \mathbf{F}_{in} &= \text{Concat}(\bar{\mathbf{F}}_3, \mathbf{F}_{\text{ctx}}), \\ \mathbf{F}_{dw} &= \text{LN}(\text{DWConv}(\mathbf{F}_{in}))), \end{aligned} \tag{3}$$

where $\text{LN}(\cdot)$ denotes layer normalization, $\text{DWConv}(\cdot)$ denotes depthwise separable convolution, and $\text{Concat}(\cdot)$ indicates channel-dimension concatenation operation. This output representation $\mathbf{F}_{dw} \in \mathbb{R}^{2C\times H\times W}$ captures both local details and global context.

For efficient computation of attention weights in dynamic kernel generation, we first split the output feature map along its channel dimension to obtain query-key pairs: $[\text{Q}, \text{K}] = \text{Split}(\mathbf{F}_{dw})$. We then reshape these tensors to matrix form: $\text{Q}, \text{K} \in \mathbb{R}^{C\times HW}$. To enable the convolution to be adaptive over spatial regions, we introduce two hyperparameters α_1 and α_2 that dynamically control the kernel size:

$$\hat{\text{K}} = \text{Linear}(\text{K}), \hat{\text{K}} \in \mathbb{R}^{C\times(\alpha_1^2+\alpha_2^2)}, \tag{4}$$

where $\hat{\text{K}}$ provides new reference patterns to compute attention weights. Subsequently, the features are projected into attention weights and two sets of dynamic kernel templates:

$$\begin{aligned} \mathbf{A} &= \text{Q}^\top \hat{\text{K}} \in \mathbb{R}^{HW\times(\alpha_1^2+\alpha_2^2)}, \\ [\mathbf{A}_1, \mathbf{A}_2] &= \text{Softmax}(\text{Split}(\mathbf{A})), \end{aligned} \tag{5}$$

where $\mathbf{A}$ encodes spatial relationships for both kernel scales. $\text{Split}(\cdot)$ performs channel-wise separation into α_1^2 and α_2^2 components. $\text{Softmax}(\cdot)$ applies independent normalization to each kernel template. $\mathbf{A}_1$ and $\mathbf{A}_2$ correspond to the dynamic weights for $\alpha_1 \times \alpha_1$ and $\alpha_2 \times \alpha_2$ kernels respectively. These weights reshape into spatial kernels $\mathbb{R}^{H\times W\times\alpha_1\times\alpha_1}$ and $\mathbb{R}^{H\times W\times\alpha_2\times\alpha_2}$ for position-wise convolution.

To achieve spatially dynamic receptive field modeling, we apply $\mathbf{A}_1$ and $\mathbf{A}_2$ to perform weighted aggregation over the neighborhood at each spatial location for F_{dw}:

$$\begin{aligned} \hat{\mathbf{F}}_{dw} &= \text{Conv}_{down}(\mathbf{F}_{dw}), \\ \mathbf{F}_{out} &= \text{Conv}_{up}(\text{BN}(\text{Concat}(\mathbf{A}_1 * \hat{\mathbf{F}}_{dw}, \mathbf{A}_2 * \hat{\mathbf{F}}_{dw}))), \end{aligned} \tag{6}$$

where $*$ denotes attention-weighted aggregation applied locally at each position. Conv_{down} reduces channel dimensions for efficiency. Conv_{up} restores the original channel dimensions. This process can be interpreted as dynamically generating position-specific convolutional kernels that adaptively aggregate local neighborhoods at two different scales. To further enhance contextual modeling depth, we adopt a two-stage cascaded CDA to enable progressive feature fusion through sequential refinement.

Dynamic Contrastive Learning (DCL)

To mitigate the interference caused by mixing hard negatives with easy negatives in conventional binary (positive/negative) contrastive learning, and reduce training bias, we introduce a Dynamic Contrastive Learning (DCL) module. As illustrated in Figure 4 (b), we remap the original identity labels into three distinct categories:

$$\text{Label}(\boldsymbol{x}_i) = \begin{cases} 1\ (pos), & \text{if } 0 < \text{ID}(\boldsymbol{x}_i) < \delta, \\ 0\ (neg), & \text{if ID}(\boldsymbol{x}_i) = 0, \\ 2\ (hard), & \text{if ID}(\boldsymbol{x}_i) = \delta, \end{cases} \tag{7}$$

where $\boldsymbol{x}_i$ denotes the identity index of the i-th instance, with pos, neg, and $hard$ representing positive, easy negative, and hard negative instances, respectively. Here, δ is a predefined threshold for filtering invalid or ambiguous identity labels (i.e., pseudo-pedestrians).

To balance the contributions of different sample categories to the loss, we dynamically adjust their weights in proportion to the inverse frequency of each category within the batch, as formalized below:

$$\begin{bmatrix} w_{\text{pos}} \\ w_{\text{neg}} \\ w_{\text{hard}} \end{bmatrix} = \begin{bmatrix} 1.0 + (|neg| + |hard|)/N \\ 1.0 + (|pos| + |hard|)/N \\ 1.0 + (|pos| + |neg|)/N \end{bmatrix}, \tag{8}$$

where $N = pos + neg + hard$ denotes the total number of samples in the batch. Here, w_{pos}, w_{neg} and w_{hard} represent the weights for the intra-class aggregation term of positive samples, the separation term of easy negative samples, and the separation term of hard negative samples, respectively. This weighting scheme assigns higher importance to scarcer sample types, thereby promoting training balance.

Based on these dynamic weights, we formulate the dynamic contrastive learning loss as:

$$\mathcal{L}_{\text{DCL}} = w_{\text{pos}} \cdot D_{\text{pos}} + w_{\text{neg}} \cdot D_{\text{neg}} + w_{\text{hard}} \cdot D_{\text{hard}}, \tag{9}$$

where D_{pos} measures the intra-class distance among positive samples to enforce feature compactness, while D_{neg} and D_{hard} quantify the distances from easy and hard negative samples to the centroid of the positive distribution, respectively. Formally, let $\boldsymbol{\mu}_p$ denote the mean embedding of positive samples in the current batch. The distance terms are defined as:

$$\begin{bmatrix} D_{\text{pos}} \\ D_{\text{neg}} \\ D_{\text{hard}} \end{bmatrix} = \begin{bmatrix} \frac{1}{|pos|} \sum\limits_{\boldsymbol{x}_i \in pos} \|\boldsymbol{x}_i - \boldsymbol{\mu}_p\|^2 \\ \frac{1}{|neg|} \sum\limits_{\boldsymbol{x}_i \in neg} \left[\max\left(0, m - \|\boldsymbol{x}_i - \boldsymbol{\mu}_p\|^2\right)\right] \\ \frac{1}{|hard|} \sum\limits_{\boldsymbol{x}_i \in hard} \left[\max\left(0, m - \|\boldsymbol{x}_i - \boldsymbol{\mu}_p\|^2\right)\right] \end{bmatrix}, \tag{10}$$

where m is the margin hyperparameter.

Optimization. During training, we employ standard classification and regression losses for person detection. For person re-identification, we follow established practices (Yu et al. 2022; Xiao et al. 2017) by combining both detection and ReID losses. Here, our main contribution is the introduction of a novel Dynamic Contrastive Leaning loss (DCL), denoted as $\mathcal{L}_{\text{DCL}}$, which optimizes feature compactness within identities and separability between identities. This formulation generates more discriminative and robust representations under challenging scenarios. The complete objective function integrates these components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Det}} + \mathcal{L}_{\text{Reid}} + \lambda \mathcal{L}_{\text{DCL}}, \tag{11}$$

where λ controls the contribution of $\mathcal{L}_{\text{DCL}}$.

Experiments

Dataset and Evaluation Protocols

We evaluate our method on three datasets, i.e., CUHK-SYSU (Xiao et al. 2017), PRW (Zheng et al. 2017) and our E-Bike Rider Search (EBRS) dataset. **CUHK-SYSU** is a large-scale person search dataset, which contains 18,184 street and movie images with a total of 96,143 bounding boxes covering 8,432 identities. Following the standard protocol, the dataset is split into 5,532 training identities (11,206 frames) and 2,900 test queries (6,978 frames), with a default gallery size of 100 images, which can be extended up to 4,000. **PRW** dataset consists of 11,816 video frames captured by 6 different cameras in a campus setting, containing 43,110 bounding boxes, of which 34,304 correspond to 932 labeled identities, and the remainder are unlabeled. The dataset is divided into 482 training identities (5,704 images) and 2,057 test queries (6,112 images). For each query, all images in the test set except the query image itself serve as the gallery. To address the lack of scenario diversity in existing benchmarks, we introduce a **EBRS** dataset dedicated to e-bike riders (see Section 3 for details). It comprises 27,501 frames (35,136 boxes, 963 identities) captured from eight synchronized cameras in urban traffic scenarios. For fair comparison, we follow the standard protocols, i.e., mean Average Precision (mAP) and top-1 accuracy. The bold numbers indicate best person search performance.

Implementation Details

We implement our model using PyTorch with ResNet50 (He et al. 2016) pre-trained on ImageNet (Deng et al. 2009) as the backbone network, initializing our baseline with SeqNet (Li and Miao 2021). The network is trained for 12 epochs using SGD optimizer with a momentum of 0.9 and an initial learning rate of 0.003, which undergoes a warm-up phase in the first epoch and decays at the 10th epoch. For the CUHK-SYSU dataset, we set the OIM cyclic queue size to 5000 with a batch size of 3. For the PRW dataset the queue size is reduced to 500 while maintaining the same batch size. During inference, we apply NMS with thresholds of 0.4 or 0.5 to prune redundant detection boxes from the first and second heads respectively. All experiments are conducted on an NVIDIA RTX A6000 GPU.

Comparison with State-of-the-Art Methods

Results on EBRS. As shown in Table 1, our method achieves the best performance on the EBRS dataset, with a mAP of 44.9% and a top-1 accuracy of 88.6%. Compared with baseline method SeqNet (Li and Miao 2021), our method improves the mAP by +15.3% and top-1 accuracy by +2.4%. Compared with SAT method (Fiaz et al. 2023) based on scale-aware transformer architecture, our method surpasses it by +5.0% in mAP and +0.6% in top-1 accuracy. Similarly, when compared to COAT method (Yu et al. 2022), which leverages an occluded attention mechanism, our method outperforms it by +4.8% in mAP and +1.3% in top-1 accuracy. These results clearly demonstrate that despite greater variations in camera viewpoints, occlusions, and illumination conditions within the EBRS dataset, our proposed CDCL method consistently delivers better accuracy and robustness.

Results on CUHK-SYSU and PRW. We also evaluate the effectiveness of our method on two well-established benchmarks in person search, i.e., CUHK-SYSU and PRW. As shown in Table 2, on CUHK-SYSU dataset (with a gallery size of 100), our method attains a mAP of 94.4% and a top-1 accuracy of 94.7%. This shows an improvement over the SeqNet baseline (Li and Miao 2021) while maintaining competitive performance with advanced end-to-end methods

Methods	mAP	top-1
NAE (Chen et al. 2020b)	19.5	59.7
NAE+ (Chen et al. 2020b)	24.6	68.4
OIMNet+ (Lee et al. 2022)	28.9	87.1
SeqNet (Li and Miao 2021)	29.6	86.2
SAT (Fiaz et al. 2023)	39.9	88.0
COAT (Yu et al. 2022)	40.1	87.3
SEAS (Jiang et al. 2024)	23.8	60.1
Ours	**44.9**	**88.6**

Table 1: Comparison with other SOTA methods on EBRS dataset (in %).

Methods	CUHK-SYSU		PRW	
	mAP	top-1	mAP	top-1
Two-stage				
IDE (Zheng et al. 2017)	–	–	20.5	48.3
IGPN (Dong et al. 2020b)	90.3	91.4	47.2	87.0
RDLR (Han et al. 2019)	93.0	94.2	42.9	70.2
TCTS (Wang et al. 2020)	93.9	95.1	46.8	87.5
End-to-end				
OIM (Xiao et al. 2017)	75.5	78.7	21.3	49.9
QEEPS (Munjal et al. 2019)	88.9	89.1	37.1	76.7
HOIM (Chen et al. 2020a)	89.7	90.8	39.8	80.4
BINet (Dong et al. 2020a)	90.0	90.7	45.3	81.7
NAE+ (Chen et al. 2020b)	92.1	92.9	44.0	81.1
AGWF (Han, Ko, and Sim 2021)	93.3	94.2	53.3	87.7
AlignPs+ (Yan et al. 2021)	94.0	94.5	46.1	82.1
SeqNet (Li and Miao 2021)	93.8	94.6	46.7	83.4
COAT (Yu et al. 2022)	94.2	94.7	53.3	87.4
OIMNet++ (Lee et al. 2022)	93.1	93.9	46.8	83.9
DMRNet++ (Han et al. 2023)	94.4	95.5	51.0	86.8
SAT (Fiaz et al. 2023)	94.4	94.8	54.5	87.5
PSTR (Cao et al. 2022)	93.5	95.0	49.5	87.8
SeqNeXt (Jaffe and Zakhor 2023)	94.1	94.7	50.8	86.0
PGSFL (Kim et al. 2021)	92.3	94.7	44.2	85.2
SEAS (Jiang et al. 2024)	**96.2**	**97.1**	52.0	85.7
PAD (Kim, Lee, and Sohn 2025)	94.6	95.4	**54.8**	**88.2**
Ours	94.4	94.7	54.0	85.8

Table 2: Comparison with other SOTA two-stage and end-to-end methods on CUHK-SYSU and PRW datasets (in %).

such as SAT (Fiaz et al. 2023), SeqNeXt (Jaffe and Zakhor 2023), and AlignPS+ (Yan et al. 2021). The PRW dataset presents a more challenging setting due to its larger gallery size and fewer training samples. In this scenario, our method achieves a competitive mAP of 54.0% and a top-1 accuracy of 85.8%. Compared with the baseline model SeqNet (Li and Miao 2021), it improves mAP by +7.3% and top-1 accuracy by +2.4%. Additionally, our method surpasses two-step methods like RDLR (Han et al. 2019) by +11.1% in mAP and +15.6% in top-1 accuracy, and outperforms end-to-end method such as COAT (Yu et al. 2022) by +0.7% in mAP. Furthermore, it delivers comparable results to state-of-the-art approaches like SAT (54.5% mAP) and PAD (54.8% mAP). These findings further underscore the robustness and generalization capabilities of our proposed CDCL in diverse and complex person search scenarios.

Methods	mAP	top-1
Baseline	48.0	83.5
Baseline + CMC (w/o CDA)	51.0	84.2
Baseline + CMC	53.3	84.5
Baseline + DCL	50.3	85.5
Baseline + CMC + DCL	**54.0**	**85.8**

Table 3: Ablation study of different components on PRW dataset, where 'w/o' indicates the removal of the component.

Qualitative Results. As illustrated Figure 5, we present the top-1 retrieval results for several query images on EBRS dataset under some challenging conditions, such as occlusion (first row), intense illumination (second row), diverse poses (third row), and significant scale variations (fourth row). The red/green boxes denote incorrect/correct matches. We can see that existing methods suffer from false positives and missed detections in these difficult scenarios. For example, SeqNet (Li and Miao 2021), SAT (Fiaz et al. 2023), and COAT (Yu et al. 2022) often misidentify visually similar but different targets or fail to localize heavily occluded individuals accurately. By contrast, our CDCL method consistently yields precise localization and accurate identification, demonstrating its robust ability to capture multi-level and multi-region context dependencies.

Ablation Study

Analysis of Different Components. In this work, we propose a novel Context-aware Dynamic Contrastive Learning network (CDCL), which mainly consists of two key modules: the Context Mixed Convolution (CMC) module and the Dynamic Contrastive Learning (DCL) module. To evaluate the effectiveness of each component, we conduct an ablation study on the PRW dataset. The detailed results are summarized in Table 3, where each components are progressively integrated into the baseline model. To fully analyze the effectiveness of the CMC module, we first add a variant of the CMC module without the Context-aware Dynamic Aggregation (CDA) component and thus obtain an improvement of +3.0% in mAP, demonstrating the effectiveness of this partial CMC design. Incorporating the full CMC module further enhances the performance by +2.3% in mAP. Furthermore, the addition of the DCL module to the baseline boosts the performance of the baseline network. Finally, combining both CMC and DCL modules results in the best performance. This synergy indicates that the two modules provide complementary benefits, where CMC enhances contextual feature extraction and DCL strengthens feature discrimination. In summary, these results validate the effectiveness of each components in the proposed CDCL.

Effect of Kernel Size. To analyze the influence of different convolution kernel sizes (i.e., α_1 and α_2 in Eqs. (4-

(a) Query (b) SeqNet (c) SAT (d) COAT (e) CDCL (Ours)

Figure 5: Qualitative comparison of our CDCL with SeqNet, SAT and COAT methods on the EBRS dataset. The yellow bounding boxes denote the queries, while the red and green bounding boxs indicate incorrect and correct top-1 matches, respectively.

α_1	α_2	mAP	top-1	ΔmAP	Δtop-1
3	5	53.6	85.4	-0.4	-0.4
5	7	**54.0**	**85.8**	**0.0**	**0.0**
7	9	53.3	85.2	-0.7	-0.6

Table 4: Comparison results of different convolution kernel sizes on PRW dataset.

5)) within the Context Mixed Convolution (CMC) module, we conduct an ablation study on the PRW dataset, as summarized in Table 4. The results indicate that our CDCL model yields the best performance when we set $\alpha_1 = 5$ and $\alpha_2 = 7$. When we further set $\alpha_1 = 7$ and $\alpha_2 = 9$, this leads to a slight decrease in the performance of our CDCL model in both mAP and top-1 accuracy. It shows that excessively large kernels introduce redundant information and increase model complexity without corresponding gains in generalization. This analysis highlights the importance of appropriately balancing the receptive field size to capture sufficient contextual information while avoiding over-parameterization in the CMC module.

Hyperparameter Analysis: The hyperparameter analysis reveals that optimal performance is achieved with a margin $m = 1.3$ in the contrastive loss, effectively balancing intra-class compactness and inter-class separability, while deviations from this value degrade results. Similarly, the contrastive loss weighting factor λ demonstrates peak effectiveness at $\lambda = 0.1$, striking an ideal trade-off between discriminative embedding learning and training stability, with performance declining for values outside this optimal range.

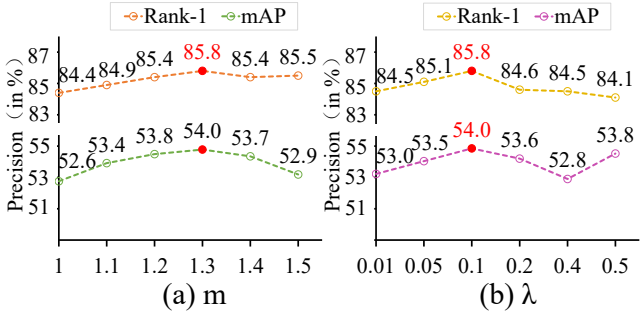

Figure 6: Performance results of different parameter settings on PRW dataset.

These findings highlight the sensitivity of model performance to precise hyperparameter tuning.

Conclusion

To the best of our knowledge, we construct the first E-Bike Rider Search (EBRS) dataset for person search in challenging urban traffic scenarios. We propose a novel Context-aware Dynamic Contrastive Learning (CDCL) framework that combines context mixed convolution and dynamic contrastive learning to robustly handle occlusions, pose variations, and scale changes. The framework further enhances discriminative power through hard sample mining strategies and adaptive loss weighting, significantly improving intra-class compactness and inter-class separability. Comprehensive experiments demonstrate superior performance under demanding urban surveillance conditions.

Acknowledgments

This research is supported in part by the National Natural Science Foundation of China (62302013, 62272006, 62472006) and the Natural Science Foundation of Anhui Province (2308085QF220).

References

Cao, J.; Pang, Y.; Anwer, R. M.; Cholakkal, H.; Xie, J.; Shah, M.; and Khan, F. S. 2022. PSTR: End-to-End One-Step Person Search with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9458–9467.

Chen, D.; Zhang, S.; Ouyang, W.; Yang, J.; and Schiele, B. 2020a. Hierarchical online instance matching for person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 10518–10525.

Chen, D.; Zhang, S.; Ouyang, W.; Yang, J.; and Tai, Y. 2018. Person search via a mask-guided two-stream CNN model. In *Proceedings of the European Computer Vision Conference*, 734–750.

Chen, D.; Zhang, S.; Yang, J.; and Schiele, B. 2020b. Norm-aware embedding for efficient person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12615–12624.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 248–255.

Dong, W.; Zhang, Z.; Song, C.; and Tan, T. 2020a. Bi-directional interaction network for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2839–2848.

Dong, W.; Zhang, Z.; Song, C.; and Tan, T. 2020b. Instance guided proposal network for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2585–2594.

Fiaz, M.; Cholakkal, H.; Anwer, R. M.; and Khan, F. S. 2023. SAT: Scale-Augmented Transformer for Person Search. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4809–4818.

Han, B.-J.; Ko, K.; and Sim, J.-Y. 2021. End-to-end trainable trident person search network using adaptive gradient propagation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 925–933.

Han, C.; Ye, J.; Zhong, Y.; Tan, X.; Zhang, C.; Gao, C.; and Sang, N. 2019. Re-id driven localization refinement for person search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9814–9823.

Han, C.; Zheng, Z.; Su, K.; Yu, D.; Yuan, Z.; Gao, C.; Sang, N.; and Yang, Y. 2023. DMRNet++: Learning Discriminative Features With Decoupled Networks and Enriched Pairs for One-Step Person Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6): 7319–7337.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Huang, W.; Jia, X.; Zhong, X.; Wang, X.; Jiang, K.; and Wang, Z. 2023. Beyond the parts: Learning coarse-to-fine adaptive alignment representation for person search. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(3): 1–19.

Jaffe, L.; and Zakhor, A. 2023. Gallery Filter Network for Person Search. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1684–1693.

Jiang, Y.; Wang, H.; Peng, J.; Fu, X.; and Wang, Y. 2024. Scene-Adaptive Person Search via Bilateral Modulations. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 938–946.

Kim, H.; Joung, S.; Kim, I.-J.; and Sohn, K. 2021. Prototype-Guided Saliency Feature Learning for Person Search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4863–4872.

Kim, H.; Lee, J.; and Sohn, K. 2025. Prototype-Guided Attention Distillation for Discriminative Person Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1): 99–115.

Lan, X.; Zhu, X.; and Gong, S. 2018. Person search by multi-scale matching. In *Proceedings of the European Conference on Computer Vision*, 536–552.

Lee, S.; Oh, Y.; Baek, D.; Lee, J.; and Ham, B. 2022. OIM-Net++: Prototypical Normalization and Localization-Aware Learning for Person Search. In *Proceedings of the European Computer Vision Conference*, volume 13670, 622–640.

Li, Z.; and Miao, D. 2021. Sequential end-to-end network for efficient person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2011–2019.

Luo, H.; Jiang, W.; Gu, Y.; Liu, F.; Liao, X.; Lai, S.; and Gu, J. 2019. A Strong Baseline and Batch Normalization Neck for Deep Person Re-identification. *IEEE Transactions on Multimedia*, 22(10): 2597–2609.

Munjal, B.; Amin, S.; Tombari, F.; and Galasso, F. 2019. Query-guided end-to-end person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 811–820.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6): 1137–1149.

Tian, Y.; Chen, D.; Liu, Y.; Yang, J.; and Zhang, S. 2024. Divide and conquer: Hybrid pre-training for person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5224–5232.

Wang, C.; Ma, B.; Chang, H.; Shan, S.; and Chen, X. 2020. Tcts: A task-consistent two-stage framework for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11952–11961.

Wang, J.; Pang, Y.; Cao, J.; Sun, H.; Shao, Z.; and Li, X. 2024. Deep intra-image contrastive learning for weakly supervised one-step person search. *Pattern Recognition*, 147: 110047.

Xiao, T.; Li, S.; Wang, B.; Lin, L.; and Wang, X. 2017. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3415–3424.

Xu, Y.; Ma, B.; Huang, R.; and Lin, L. 2014. Person search in a scene by jointly modeling people commonness and person uniqueness. In *Proceedings of the ACM International Conference Multimedia*, 937–940.

Yan, Y.; Li, J.; Qin, J.; Bai, S.; Liao, S.; Liu, L.; Zhu, F.; and Shao, L. 2021. Anchor-free person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7690–7699.

Yu, R.; Du, D.; LaLonde, R.; Davila, D.; Funk, C.; Hoogs, A.; and Clipp, B. 2022. Cascade transformers for end-to-end person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7267–7276.

Zheng, L.; Zhang, H.; Sun, S.; Chandraker, M.; Yang, Y.; and Tian, Q. 2017. Person re-identification in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1367–1376.