Multi-Granularity Graph-Convolution-Based Method for Weakly Supervised Person Search

Haichun Tai¹, De Cheng¹*, Jie Li¹, Nannan Wang¹* and Xinbo Gao²
¹Xidian University
²Chongqing University of Posts and Telecommunications
htai@stu.xidian.edu.cn, {dcheng, leejie, nnwang}@xidian.edu.cn, gaoxb@cqupt.edu.cn

Abstract

One-step Weakly Supervised Person Search (WSPS) jointly performs pedestrian detection and person Re-IDentification (ReID) only with bounding box annotations, which makes the traditional person ReID problem more practical and efficient for real-world applications. However, this task is very challenging due to the following reasons: 1) large feature gap between person ReID and general object detection tasks when learning shared representations; 2) difficult pseudo identity estimation for each person image with unrefined raw detection and dramatic scale changes. To address the above issues, we propose a multi-granularity graph convolution framework to jointly optimize the aligned task features, as well as to assist the pseudo label estimation. Specifically, the multi-granularity feature alignment module (MFA) in the designed two-branch framework employs cluster-level bi-directional interaction of various granularity information to narrow down the large feature gap. Further, upon the MFA module, we introduce the multi-granularity neighbor-guided graph-convolution-based pseudo-label estimation module, to enhance feature representations for distinguishing diverse identities. Extensive experimental results demonstrate the effectiveness of the proposed method, and show superior performances to state-of-the-art methods by a large margin on the CUHK-SYSU and PRW datasets.

1 Introduction

Person Search (PS) aims to localize a given query person in a gallery of unconstrained scene images, which is a joint task of pedestrian detection and ReID [Dong et al., 2020a; Li and Miao, 2021; Yu et al., 2022]. It makes the traditional person ReID problem more practical and efficient for real-world applications, such as video surveillance, retail analytics, smart cities, etc. However, PS is more challenging due to the unrefined raw detection, dramatic scale and camera view changes, etc. Most existing PS methods adopt fully-supervised settings, termed SPS [Cao et al., 2022; Li and Miao, 2021], where both of the bounding boxes and identity labels are required for model training. Although such SPS methods have achieved high performance, it is very expensive and time-consuming to annotate all bounding boxes and identities, especially for obtaining a large number of different identity annotations across different scenes. Considering that it is much easier to annotate bounding boxes than person identities, our work focuses on WSPS, where only the bounding-box annotations are required for model training.

Existing WSPS methods can be roughly divided into two categories: the two-step [Jia et al., 2022] and one-step [Wang et al., 2023; Wang et al., 2024] WSPS methods. The two-step WSPS methods first train a detector to crop person images and then apply an unsupervised learning (USL) ReID method for person retrieval. Though achieving higher performance, they suffer low efficiency, requiring high computation costs with two network parameters during training and being inconvenient for testing. In contrast, the one-step WSPS method jointly learns the person detector and identifier, which is more efficient and practical to be seamlessly integrated into real-world applications, with improved speed and resource utilization. Therefore, our paper focuses on the one-step WSPS.

For the one-step WSPS, it faces two main challenges: The first is the feature gap between person ReID and general object detection tasks [Han et al., 2021b; Dong et al., 2020a], where PS is a fine-grained task requiring accurate human appearance information, which is more likely to be affected by some redundant context information outside the bounding box. On the contrary, the general object detector usually needs to be robust to some subtle appearance changes. Such a feature gap is also the reason for the great performance gap between one-step and two-step SPS tasks. Secondly, how to accurately estimate the pseudo-identity for each person image is another challenge for WSPS. Compared to traditional USL ReID, the person features in WSPS come from the low-quality proposals generated by unrefined raw detection, resulting in inferior discriminative representation. In addition, the person images in various scenes exhibit dramatic size variations, making person feature matching difficult.

To address the above issues in one-step WSPS, we propose a multi-granularity graph convolution framework to jointly optimize the aligned task features, as well as to assist the pseudo label estimation for each person image. Specifically,
for the Multi-granularity Feature Alignment (MFA) module, we aim to reduce the feature gap through task-aware information interaction, encouraging a closer feature representation between two tasks. We maintain two branches: Search branch and Instance branch. The search branch takes the whole scene image as input, and a detector is applied to obtain the detection features. The instance branch utilizes the ground truth bounding box as its input. Given diverse inputs, two branches contain different granularity information: the search branch incorporates broader scene context information, while the instance branch is expected to capture more detailed and fine-grained appearance information. Consequently, we build a multi-granularity memory bank to store the granularity information contained in these two branch features. To narrow down the feature gap, we employ Cluster-level Multi-granularity Alignment (CMA) to facilitate the cluster-level bi-directional interaction of various granularity information within the same pseudo identity. Besides, we add noise to the initial bounding boxes, obtaining additional augmented drifted boxes for feature generation. Then, we use Box-drift Invariance Learning (BIL) to enhance the search features against low-quality detection bounding boxes.

Building upon the multi-granularity feature alignment module, we further introduce the Multi-granularity Neighbor-guided Graph Convolution Based Pseudo-label Estimation (MGPE) module, to enhance feature representations for distinguishing diverse identities. We argue that the effective utilization of multi-granularity features is crucial for representing a person image. Single-granularity features lack sufficient information, while simply aligning multi-granularity features within a person image usually struggles to represent the person image effectively. To address this issue, we employ an Offline Graph Convolution (OGC) formulation for information propagation, progressively gathering favorable granularity information from adjacent nodes and then enriching the feature representation for distinguishing diverse identities. Specifically, we propose the Neighbor Guided Sparse Graph Construction (NGSGC) strategy, ensuring that reliable and valuable neighbors are used for effective granularity information propagation. Finally, following representative one-step WSPS methods [Yan et al., 2022b; Wang et al., 2024], our method also alternates between feature representation learning and pseudo-label estimation.

Our contributions are summarized as follows:

- We propose the multi-granularity feature alignment module to facilitate bi-interaction across different granularities, narrowing down the feature gap that exists in the joint training of pedestrian detection and ReID tasks for one-step WSPS.

- We introduce the multi-granularity graph-convolution-based pseudo-label estimation module, to enhance feature representations for distinguishing diverse identities, and finally obtain accurate pseudo labels for WSPS.

- Extensive experimental results demonstrate the effectiveness of the proposed method, and show superior performances to state-of-the-art methods by a large margin on the CUHK-SYSU and PRW datasets.

2 Related Work

2.1 Supervised Person Search

SPS is a task that combines pedestrian detection and person ReID with bounding boxes and person identity annotations for model training [Ke et al., 2022; Yang et al., 2021]. It can be divided into two-step and one-step SPS methods. Two-step SPS methods perform detection and Re-ID tasks, respectively [Chen et al., 2018; Dong et al., 2020b]. One representative work [Wang et al., 2020] solves the consistency problem existed between two tasks by introducing an identity-guided query detector and detection results adapted ReID model. One-step SPS methods jointly perform detection and ReID tasks with shared feature representations [Cao et al., 2022; Xiao et al., 2017; Yu et al., 2022]. [Xiao et al., 2017] first combines the two-step SPS method into a unified framework and proposes online instance matching to effectively train the ReID task. [Li and Miao, 2021] employs the sequential detection method to solve the problem of inferior features caused by low-quality proposals. Benefiting from DETR, [Cao et al., 2022] first proposes the transformer-based person search framework, achieving better re-id performance with a multi-level supervision scheme. Two-step SPS methods can achieve better performance, but they suffer from low efficiency and high computation costs. Meanwhile, the feature gap in general detection and ReID tasks poses a new challenge for the joint training of one-stage SPS method.

2.2 Weakly Supervised Person Search

WSPS only relies on easily obtained bounding boxes as annotations for the joint training of detection and ReID tasks. Similar to SPS, WSPS can also be divided into two-step and one-step methods. For two-step WSPS methods, [Jia et al., 2022] first performs pedestrian detection and then alternate between context-informed clustering and unpaired-assisted memory training. For one-step WSPS methods, [Yan et al., 2022b] and [Han et al., 2021b] solve the task by clustering the features of proposals to generate pseudo-labels and then guiding the training process. [Wang et al., 2023] proposes scale-invariant learning to solve the scale variation obstacle, and a dynamic multi-label prediction is used for progressive true label seekness. In WSPS, the generation of pseudo-labels is essential to guide the training process, and the quality of these pseudo-labels significantly affects performances. Consequently, generating accurate pseudo-labels is another crucial challenge for WSPS.

2.3 Graph Convolutional Networks.

The Graph Convolutional Network (GCN) is widely used to propagate information and measure similarity relationships among samples [Kipf and Welling, 2016]. [Zhang et al., 2020; Zhong et al., 2019] use GCN to solve the problem of noisy labels. [Yan et al., 2022a] uses GCN to correct the initial noisy pseudo-labels. However, these GCN-based methods require training. [Zhang et al., 2023b] proposes a graph convolution re-ranking method based on vector generation formulation, which allows graph information to propagate in an offline manner. Motivated by this, we introduce a neighbor-guided graph convolution strategy to enhance feature representation and finally achieve better performance.
Figure 1: This is the proposed Multi-Granularity Graph-Convolution-Based framework for one-step weakly supervised person search, which consists of two components: Multi-Granularity Feature Alignment (MFA) Representation Learning and Multi-Granularity Graph-Convolution-Based Pseudo-Label Estimation (MGPE). In the MFA stage, we narrow the large feature gap through cluster-level bi-directional interaction of various granularity information and box-drift invariance learning. In the MGPE stage, we progressively incorporate neighborhood reliable granularity information, enhancing the feature representations for distinguishing diverse identities.

3 Proposed Method

3.1 Overview

One-step WSPS jointly performs pedestrian detection and person ReID, with the bounding boxes as the only available annotations. As illustrated in Figure 1, our framework consists of two components: 1) Multi-Granularity Feature Alignment (MFA) representation learning module, which employs cluster-level bi-directional interaction of variant granularity information, to narrow down the feature gap between person ReID and object detection tasks; 2) Multi-Granularity Graph-Convolution-Based Pseudo-Label Estimation module, aiming to obtain enhanced feature representations facilitating accurate pseudo-label estimation for each person image. Intuitively, our method alternates between feature learning and pseudo-label estimation.

In MFA learning, we first build a multi-granularity memory bank to store the two branch features. For each branch, we employ intra-InfoNCE loss to assist the learning of branch-specific granularity information [Oord et al., 2018]. To facilitate interaction between distinct granularities, we introduce CMA with inter-InfoNCE loss to learn shared feature representations. Moreover, BIL enhances search branch features robust to low-quality proposals. For MGPE, we first utilize multi-granularity memory banks to construct the neighborhood-guided graph. Then, the OGC is applied to progressively utilize neighborhood granularity information, enhancing feature representation for more accurate pseudo-label estimation. Focusing on one-step WSPS, our method combines ReID loss and detection loss, like Faster R-CNN [Ren et al., 2015] for joint task training.

3.2 Multi-granularity Feature Alignment Representation Learning

We adopt a two-stream network consisting of a search branch and an instance branch with shared network architecture and parameters. Specifically, the shared architecture includes the backbone network, region proposal network (RPN) and RoI-Align layers, where RPN is only used for detection while RoI-Align layers are used for both detection and ReID feature generation. Following DICL [Wang et al., 2024], we utilize the fused feature map from the last two layers of the backbone as input for both the RPN and RoI-Align layers. For a given scene image \( I \), the search branch first uses RPN to detect the proposal bounding boxes corresponding to the \( i \)-th person image. RoI-Align layers are then employed to perform classification, regression, and feature extraction for search features \( f^S_i \in \mathbb{R}^d \) based on proposals, where \( d \) is the dimension of features. While for instance branch, the shared network takes the \( i \)-th person patch cropped with the ground truth bounding box \( b_i \) to extract the instance features \( f^I_i \in \mathbb{R}^d \).

Multi-granularity Memory Banks. To obtain the estimation of pseudo-labels and supervised information for each person image, we maintain two memory banks: search memory bank \( \mathcal{M}^S \in \mathbb{R}^{N \times d} \) and instance memory bank \( \mathcal{M}^I \in \mathbb{R}^{N \times d} \), where \( N \) is the number of person images annotated with ground truth bounding boxes. Two memory banks, termed as multi-granularity memory banks, store the features extracted from their respective branches, and the \( i \)-th feature in the memory banks is updated as follows:

\[
\mathcal{M}^*_i = m\mathcal{M}^*_i + (1-m)f^S_i,
\]

where \( * \in \{S, I\} \) denotes the item belongs to the search branch or instance branch, and \( m \) is the momentum updating.
factor. The Exponential Moving Average (EMA) updating strategy improves the robustness of features stored in memory banks [Hansu et al., 2013]. During training, each feature \( f_i \) will obtain a pseudo-label \( y_i \) through clustering. Then, the centroid \( C_k \in \mathbb{R}^d \) for the cluster \( k \) can be computed as follows:

\[
C_k^* = \frac{1}{N_k} \sum_{y_i=k} \mathcal{M}_i^*,
\]

where \( N_k \) is the number of examples contained in cluster \( k \). Specifically, we design the cluster level multi-granularity alignment inter-InfoNCE loss as follows:

\[
L_{\text{interS}} = -\frac{1}{P_S} \sum_{i=1}^{P_S} \log \frac{\exp(f_i^S, C_k^*/\tau_1)}{\sum_{k=1}^{K} \exp(f_i^S, C_k^*/\tau_1)},
\]

\[
L_{\text{interI}} = -\frac{1}{P_I} \sum_{i=1}^{P_I} \log \frac{\exp(f_i^I, C_k^*/\tau_1)}{\sum_{k=1}^{K} \exp(f_i^I, C_k^*/\tau_1)}.
\]

The corresponding overall cluster-level multi-granularity alignment inter-InfoNCE loss is as follows:

\[
L_{\text{inter}} = L_{\text{interS}} + L_{\text{interI}}.
\]

Box-drift Invariance Learning. As search features suffer from low-quality proposals and insufficient fine-grained information, we propose BIL to address the problems. Specifically, we duplicate the ground truth bounding boxes to generate \( Z \) additional boxes and then introduce random noise to them as follows:

\[
(a_1, a_2) = (a_1, a_2) + U(-\varepsilon, \varepsilon) \times (a_2 - a_1),
\]

\[
(b_1', b_2') = (b_1, b_2) + U(-\varepsilon, \varepsilon) \times (b_2 - b_1),
\]

where \((a_1, b_1)\) and \((a_2, b_2)\) represent the coordinates of the upper left corner and lower right corner of the ground truth bounding box, \( U(-\varepsilon, \varepsilon) \) denotes a random number obtained from a uniform distribution on the interval \([-\varepsilon, \varepsilon]\), \((a_1', b_1')\) and \((a_2', b_2')\) are the new augmented box-drift bounding box coordinates. Before they crop the person patch, we first mask the background pixels outside the ground truth bounding box \( b_i \), to avoid instance feature capturing redundant ground truth information. Then, we extract the augmented instance features for the \( i \)-th person image \( f_i^A = \{ f_i^1, f_i^1', f_i^2, f_i^2' \} \) corresponding to no-drift-box, drift-box-1, and drift-box-2. The box-drift invariance loss is as follows:

\[
L_{\text{box}} = \frac{1}{P_S} \sum_{i=1}^{P_S} \left( \max_{f \in f_i^A} d(f_i^S, f) - \min_{f \in f_i^A} d(f_i^S, f) + \delta \right)_+ + d(f_i^S, f^{10}),
\]

where \([ \cdot \]_+ is \( \max(\cdot, 0) \), \((\cdot, \cdot)\) is the Euclidean distance between examples, and \( \delta \) is a margin parameter, \( f_i^A = \cup_{y_i=y_i, f_i^A} \) indicates the positive examples that have the same pseudo labels with \( i \)-th person image, and \( f_i^A = \cup_{y_i\neq y_i, f_i^A} \) indicates the negative examples that have the different pseudo-labels. \( d(f_i^S, f^{10}) \) is a regularization term, avoiding missing fine-grained information caused by box drift. \( L_{\text{box}} \) searches the hard positive and hard negative examples to assist search features in learning robust features against low-quality proposals and fine-grained information aligned with instance features. Furthermore, \( f_i^A \) is utilized to update the \( i \)-th feature \( M_i^f \) in the instance memory for fine-grained information ensemble, as follows:

\[
M_i^f = M_i^f + (1 - m) \sum_{f_j \in f_i^A} w_{ij} f_j,
\]

where \( w_{ij} \) is a weight factor representing the contribution of the box-drift features \( f_j \in f_i^A \) to the no-drift feature \( f_i^{10} \):

\[
w_{ij} = \frac{\exp(f_j, f_i^{10}/\tau_2)}{\sum_{f \in f_i^A} \exp(f, f_i^{10}/\tau_2)},
\]

where \( \tau_2 \) is a temperature hyper-parameter. Masking the background pixel information allows box-drift features find more appearance details, leading to diverse fine-grained features in \( f_i^A \). The updating strategy successfully integrates various fine-grained information into a unified feature representation, enhancing the discriminative ability of features stored in the instance memory bank.

### 3.3 Multi-granularity Graph-convolution-based Pseudo-label Estimation

**Multi-granularity Graph Construction.** To thoroughly investigate the relationships among various granularity features, we construct a multi-granularity graph containing \( 2N \) nodes based on our multi-granularity memory banks. The definition of the multi-granularity graph is as follows:

\[
M = [M^S; M^I],
\]

where \([ \cdot; \cdot \] means feature concatenation in the column dimension, and \( M \in \mathbb{R}^{2N \times d} \). The initial edges of our graph are constructed by a dense affinity matrix as follows:

\[
A = M^T M,
\]
where $A \in \mathbb{R}^{2N \times 2N}$ is the cosine similarity matrix, and $A_{ij} = M_i^j$, $M_i = M[i, :]$, representing the relationship between node $i$ and node $j$. Such a dense affinity matrix suffers from low efficiency and high computation costs.

**Neighbor-Guided Sparse Graph Construction.** To mitigate the problem caused by the dense affinity matrix, we introduce the NGSGC strategy. It dynamically searches neighbors for each example, aiming to create a sparse affinity matrix from the dense matrix. To select valuable neighbors, we introduce our multi-granularity neighbor consistency hypothesis, which supposes that reliable neighbors should be simultaneously close to both granularity features within the same person image. The neighbors for the $i$-th example can be obtained as follows:

$$\mathcal{N}_i = \{ j \mid (A_{ji} \geq A_{ii}) \wedge (A_{ji} \geq A_{ii} - \beta), j \in \Gamma(A_{i,i}, k_1) \},$$

with

$$i = \begin{cases} i + N, & \text{if } i < N \\ i - N, & \text{if } i \geq N \end{cases},$$

where $\Gamma(\cdot, k_1)$ is the top-$k_1$ neighbors, ‘$\wedge$’ represents logical-and operator. Indices $1$ to $N$ represent the search feature nodes, whereas $N + 1$ to $2N$ correspond to the instance feature nodes. Obviously, $i$ and $i$ represent the indexes of two different granularity features under the same person image. $A_{ij}$ represents the similarity relationship between two granularity features. $A_{ji}$ and $A_{ji}$ denote the similarity of the $j$-th feature to the different granularity features within the $i$-th person image. $\beta$ is a relaxation coefficient used to control the selection of more reliable samples.

We use the $A_{ji}$ to select neighbors, and we treat it as a discriminative ability score for $i$-th example. A high score of $A_{ij}$ indicates that the person patch can yield similar feature representations under detection and ReID tasks in WPS. We believe that such examples already have sufficient discriminative ability to generate correct pseudo-label estimation, and there is no need for additional excessive neighbors to enrich their own feature representation. On the contrary, we consider the examples with a low score require more valuable granularity information to enhance their representations. Consequently, $A_{ij} \geq A_{ji}$ denotes that we limit the number of neighbors with different $A_{ij}$ values corresponding to different discriminative abilities. $A_{ij} \geq A_{ji} - \beta$ denotes that we further select the reliable neighbors for information propagation based on our hypothesis.

Consequently, the new sparse affinity matrix used for our OCN information propagation formulation is as follows:

$$\tilde{A}_{ij} = \begin{cases} \exp(A_{ij} / \gamma), & j \in \mathcal{N}_i \\ 0, & \text{otherwise} \end{cases},$$

where $\gamma$ is the temperature parameter. Considering that the similarity relationships between examples are pairwise, the affinity matrix should be symmetric. To achieve this, we adopt a simple averaging strategy as follows:

$$\bar{A} = (\tilde{A} + \tilde{A}^\top) / 2.$$  

**Offline Graph Convolution Feature Propagation.** The standard GCN layer can be formulated as follows:

$$\hat{X} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}} X W,$$  

where $A$ is the affinity matrix, and $D$ is the diagonal degree matrix computed with $D_{ii} = \sum_j A_{ij}$. $X$ is the input features, and $W$ is the parameter of GCN that is being trained. When $W$ is an identity matrix, the GCN is a feature propagation on the graph characterized by an affinity matrix.

Inspired by Eq. 19, our Offline Graph Convolution formulation can be rewritten as:

$$[\bar{M}^2; \bar{M}^1] = \bar{M} = D^{-\frac{1}{2}} \bar{A} D^{-\frac{1}{2}} M,$$  

where $M \in \mathbb{R}^{2N \times d}$ is the input features (worked as graph nodes). $\bar{A} \in \mathbb{R}^{2N \times 2N}$ is the sparse affinity matrix obtained by Eq. 15 and 18. The degree matrix can be recomputed with $D_{ii} = \sum_{j=1}^{2N} A_{ij}$. $\bar{M}$ is the enhanced output features, which has the same dimension as $M$. We perform such feature propagation several times, empirically setting it to 3 in our experiments. According to the experimental results, we use the features $\bar{M}^3$ for clustering and then reassigned the pseudo-labels to the features stored in memory banks and training examples.

The similarity matrix $A$ may become very large due to the necessity of considering relationships among $2N$ nodes, requiring high computational costs. To enhance computational efficiency, we could divide the similarity matrix $A$ into subgraphs and then utilize the divided subgraphs and their relationships for propagating information between nodes.

### 3.4 Training and Inference

On the whole, our model is trained in an end-to-end manner, and the overall loss arrives at:

$$L = \lambda (L_{\text{intra}} + L_{\text{inter}} + L_{\text{box}}) + L_{\text{det}},$$

where $L_{\text{det}}$ is the detection loss used in Faster R-CNN [Ren et al., 2015], and $\lambda$ is a hyper-parameter to balance the ReID loss and detection loss.

During inference, we only use the search branch to detect the person and generate the ReID features for evaluation.

### 4 Experiments

#### 4.1 Datasets and Settings

**CUHK-SYSU**[Xiao et al., 2017] is a large-scale public dataset for person search, consisting of 12,490 street images shot by cameras and 5,694 movie snapshots. It contains 18,184 images, 8,432 annotated identities, and 43,110 annotated bounding boxes in total. The training set has 11,206 images with 5,532 unique identities, while the test set consists of 6,112 images, 25,062 pedestrian bounding boxes, and 482 unique person identities. The training set has 11,206 images with 5,532 unique identities, while the test set consists of 6,112 images, 25,062 pedestrian bounding boxes, and 450 person identities.

**Evaluation Protocol.** Following the evaluation settings widely used in PS, we use the Cumulative Matching Characteristic (CMC) and the mean Averaged Precision (mAP) to evaluate the performance of our methods.

---

1330
4.2 Implementation Details

Our method is implemented on the open-source library mmdetector [Chen et al., 2019], where the backbone is built using ResNet50, pre-trained on ImageNet [He et al., 2016]. Following DICL [Wang et al., 2024], we employ the fused feature map of the last two layers of backbone for RPN and deformable layers for RoI-Align layers. For the instance branch, we crop and resize the person patch into a fixed size of \(220 \times 96\). We adopt the SGD optimizer with the batch size set to 4, the epochs set to 26, the momentum set to 0.9, and the weight decay set to 0.0005. The initial learning rate is 0.003, which is divided by 10 at 22\textsuperscript{nd} epochs. For all gallery sizes, our method consistently outperforms existing weakly supervised methods, demonstrating its superior performance approximation to the two-stage method. In comparison to some SPS methods, our approach surprisingly outperforms them.

### 4.3 Comparison with the State-of-the-arts

In this section, we compare our method with current state-of-the-art methods, including fully supervised methods and weakly supervised methods. All methods can be divided into one-step and two-step methods.

**Performance on PRW.** In Table 1, we compare our method with existing WSPS methods. PRW poses a greater challenge with a larger gallery size (6112) compared to CUHK-SYSU (default 100). For the one-step WSPS, our method significantly outperforms all existing methods, surpassing the state-of-the-art method DICL mAP/top-1 by 4.2%/1.5%. Though a performance gap still exists for two-step WSPS, our method provides a more powerful performance approximation to the two-stage method. In comparison to some SPS methods, our approach surprisingly outperforms them.

**Performance on CUHK-SYSU.** Table 1 displays the mAP/top-1 results on the CUHK-SYSU dataset using the default gallery size of 100. Unlike PRW, which has a constant gallery size, CUHK-SYSU provides multiple gallery sizes ranging from 50 to 4,000. To assess the scalability of our method, we compare it with existing one-step WSPS methods across different gallery sizes, as shown in Fig. 2. Across all gallery sizes, our method consistently outperforms existing one-step WSPS methods, demonstrating its superior capabilities in handling large and challenging datasets.

### 4.4 Ablation Study

In our proposed method, the framework is built on a two-stream network, containing a search branch and an instance branch. In Table 2, according to Eq. 21, the “base” denotes our baseline method, which contains the items \(\lambda \mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{det}}\) for model training, where \(\mathcal{L}_{\text{intra}}\) is the ReID intra-InfoNCE and \(\mathcal{L}_{\text{det}}\) is the detection loss. On top of the baseline, there are three components in our proposed method: 1) the Inter-

<table>
<thead>
<tr>
<th>base</th>
<th>(\mathcal{L}_{\text{inter}})</th>
<th>(\mathcal{L}_{\text{box}})</th>
<th>MGPE</th>
<th>PRW mAP</th>
<th>PRW top-1</th>
<th>CUHK-SYSU mAP</th>
<th>CUHK-SYSU top-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>28.0</td>
<td>76.6</td>
<td>84.1</td>
<td>85.9</td>
</tr>
<tr>
<td>2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>32.0</td>
<td>79.3</td>
<td>87.0</td>
<td>88.7</td>
</tr>
<tr>
<td>3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>30.5</td>
<td>78.5</td>
<td>86.3</td>
<td>88.1</td>
</tr>
<tr>
<td>4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>34.2</td>
<td>80.6</td>
<td>87.2</td>
<td>88.9</td>
</tr>
<tr>
<td>5</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>33.8</td>
<td>80.6</td>
<td>87.4</td>
<td>89.1</td>
</tr>
<tr>
<td>6</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>36.5</td>
<td>81.3</td>
<td>88.2</td>
<td>89.7</td>
</tr>
<tr>
<td>7</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>36.0</td>
<td>80.9</td>
<td>88.1</td>
<td>89.4</td>
</tr>
<tr>
<td>8</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>37.6</td>
<td>82.0</td>
<td>88.4</td>
<td>89.9</td>
</tr>
</tbody>
</table>
InfoNCE objective $\mathcal{L}_{inter}$, as mentioned in section 3.2; 2) the box-drift invariance learning $\mathcal{L}_{box}$ with fine-grained feature ensemble updating strategy, as mentioned in section 3.2; 3) the multi-granularity graph-based pseudo-label estimation method “MGPE”, as mentioned in section 3.3.

To reveal how each component contributes to the performance improvement, we implement eight variants of our method as shown in Table 2, where the check mark indicates the application of the component. Each row in Table 2 corresponds to one variant of our method, where we have completed the ablation experiment on both PRW and CUHK-SYSU datasets. For $\mathcal{L}_{inter}$, a large enhancement in mAP/top-1 performance is observed by comparing rows 8, 6, 5, and 2 with rows 7, 4, 3, and 1, respectively, with an average improvement of 2.8%/1.7% on PRW and 1.6%/1.0% on CUHK-SYSU. Using a similar comparison method, additional mAP/top-1 improvements of 1.8%/0.8% and 0.9%/0.8% are observed for $\mathcal{L}_{box}$. As for MGPE, a more substantial mAP/top-1 improvement averaging 4.9%/2.5% and 1.6%/1.5% is achieved.

**Further Illustration of Cluster-level Multi-granularity Alignment.** To further explain the effectiveness of component $\mathcal{L}_{inter}$, we investigate instance-level multi-granularity feature alignment and unidirectional information interaction. In Table 3, $\lambda \mathcal{L}_{intra} + \mathcal{L}_{det}$ serves as our baseline. $+\mathcal{L}_{inter}$ (Batch instance) indicates that we replace the $K$ cluster-level centroids in Eq. 7 with $P^*$ instance-level person features in the current mini-batch, where $P^*$ person features are assigned labels with unique batch feature indexes. Taking Eq. 5 as an example, we replace $C_{i_2}^*$ and $C_{i_2}^j$ with $f_{i_2}^*$ and $f_{i_2}^j$, respectively, where $f_i^* = f_i^j$ and $j \in 1, ..., P^*$. While $+\mathcal{L}_{inter}$ (Memory instance) indicates we replace the centroids in Eq. 7 with $N$ instance-level person features in the memory bank. Similarly in Eq. 5, $C_{i_2}^*$ and $C_{i_2}^j$ are replaced with $M_i^*$ and $M_i^j$, where $M_i^* = M_i^j$ and $j \in 1, ..., N$. The results demonstrate that our cluster-level feature alignment method $+\lambda \mathcal{L}_{inter}(L^S_{inter} + L^j_{inter})$ outperforms instance-level feature alignment and unidirectional information interaction, indicating the superior effectiveness of cluster-level and bi-interaction strategies in our method.

**Further Illustration of Multi-granularity Graph-based Pseudo-label Estimation.** To investigate the effectiveness of enhanced features for pseudo-label estimation, we conduct comprehensive experiments on the dataset PRW with various features shown in Table 4. $\tilde{M}^S$ and $\tilde{M}^I$ mean directly using the features stored in two memory banks for the generation of pseudo-labels. $\tilde{M}^S$ and $\tilde{M}^I$ denote using our OGC to enhance the feature representations and $\text{Mean}(f_1, f_2) = (f_1 + f_2)/2$. As shown in Table 4, enhanced search features $\tilde{M}^S$ achieve the best performance, which suggests that context information within search features is more critical than fine-grained appearance information contained in instance branch for pseudo-label estimation. We infer that during the inference stage, the features used for evaluation are obtained from the search branch. Therefore, emphasizing context information in pseudo label estimation is more beneficial for enhancing model performance. Furthermore, $\tilde{M}^S w/ \text{top-k}_1$ neighbors OGC means directly utilizing the raw top-$\text{top-k}_1$ neighbors to propagate information. The inferior performance illustrates the effectiveness of our neighbor selection method.

**Visualization Analysis.** Fig. 3 gives some qualitative results of our method on PRW and CUHK-SYSU test sets.

**5 Conclusion**

This paper proposes a multi-granularity graph-convolution-based framework for one-step weakly supervised person search. Facing the challenges of a large feature gap and difficult pseudo-identity estimation, we introduce the multi-granularity feature alignment and multi-granularity graph-convolution-based pseudo-label estimation modules, to narrow down the feature gap and enhance feature representation for distinguishing diverse identities, respectively. Extensive experiments on the CUHK-SYSU and PRW datasets demonstrate the superiority of our method.
Acknowledgments
This work was supported in part by the National Key R&D Program of China under Grant No.2023YFA1008600, in part by NSFC under Grant NO.62176198, U22A2096 and 62106184, in part by the Key R&D Program of Shaanxi Province under Grant 2024GX-YBXM135, in part by the Key Laboratory of Big Data Intelligent Computing under Grant BDIC-2023-A-004.

References


