Optimizing Label Assignment for Weakly Supervised Person Search

Haiyang Zhu, Xi Yang*, Nannan Wang

State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, China

zhuhaiyang@stu.xidian.edu.cn, yangx@xidian.edu.cn, nnwang@xidian.edu.cn

Abstract

Weakly supervised person search aims to detect and match individuals jointly using only bounding box annotations. The existing methods primarily employ clustering to generate pseudo labels and then proceed to training, alternating between these two stages. In the clustering stage, label assignment tasks are handled at the instance level, while in the training stage, they are managed at the proposal level. In the clustering phase, the conventional use of the DBSCAN algorithm for clustering pedestrian instance features often neglects key contextual information such as scene context and relative positioning of individuals. During the training phase, the Region Proposal Network assigns labels based on the MaxIoU, which tends to produce locally ambiguous labels. Finally, the proposals updated to the memory bank with extensive background information tend to interfere with the task of pseudo-label generation. To address these issues, this paper proposes an Optimizing Label Assignment (OLA) for weakly supervised person search. Firstly, in the clustering phase, *Context Aware Clustering* is introduced to integrate contextual information and constraints, enhancing the accuracy of clustering. Secondly, in the training phase, we adopt *Prototype Matching* based on the Optimal Transport theory to optimize label distribution from a global perspective. Furthermore, we propose *Dual Memory Bank Enhancement* that effectively enhances the accuracy of label assignment. Extensive experiments conducted on the CUHK-SYSU and PRW datasets demonstrate that our method achieves state-of-the-art performance in weakly supervised person search.

Introduction

Person search has been a challenging computer vision task, which aims to localize (detection) and re-identification (ReID) a target person from a gallery set of uncropped scene images. Successfully addressing this task has profound implications across various domains, enhancing intelligent surveillance with real-time tracking capabilities, refining behavior analysis with more accurate individual identification, and bolstering public security systems through improved crowd monitoring and management.

Recent progress in the fields of person detection (Chen et al. 2019a; Zhu et al. 2020; Sun et al. 2021) and ReID (He et al. 2021; Li et al. 2021) has significantly advanced the capabilities of person search systems. To achieve high performance, existing methods (Zheng et al. 2017; Chen et al. 2018; Xiao et al. 2017; Chen et al. 2020b; Lee et al. 2022; Yan et al. 2021) are commonly trained in a fully supervised setting where the bounding boxes and identity labels are required. However, annotating bounding boxes and identity labels on a large-scale dataset is exceptionally time-consuming and labor-intensive. Given the relative ease of annotating bounding boxes over personal identities, coupled with scale person search solutions without extensive resource investment, this paper explores weakly supervised person search that relies solely on bounding box annotations.

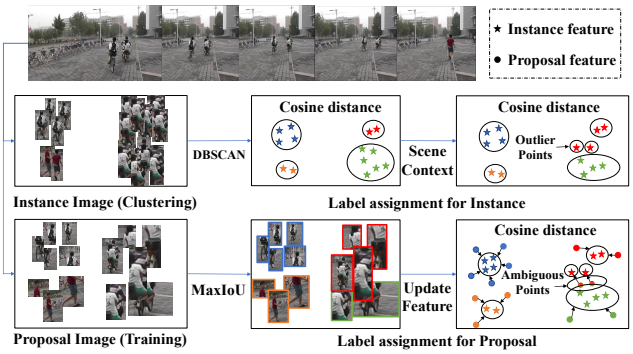


Figure 1: Label assignment in weakly supervised person search.

In weakly supervised person search, several contrastive learning based one-step methods (Yan et al. 2022; Han et al. 2021a; Wang et al. 2023, 2024) have been proposed. These methods all employ two distinct contrastive learning strategies: *memory-based contrast*, which leverages a cluster-level ReID memory bank for feature learning across the dataset, and *intra-image contrast*, which leverages contrast information within individual images for feature learning by extracting two ground truth ReID features in separate branches and enforcing their consistency both with and without contextual information. However, the above methods only focus on representation learning and ignore the issue of label inaccuracy. This challenge significantly impacts the stability of the ob-

*Corresponding author.

jective function, making it difficult to optimize. As shown in Fig. 1, in the clustering stage, the direct use of the DBSCAN algorithm predicts labels based on individual pedestrian instances, ignoring the dynamic relationships between individuals in the scene and necessary constraints (there will not be two identical pedestrians in the same scene image). Although existing methods (Yan et al. 2022; Han et al. 2021a) have considered this situation, they only filter it as outliers in the post-processing stage. During the training phase, proposals generated by the Region Proposal Network (RPN) are assigned labels using MaxIoU, which is a local matching strategy and ignores the distance relationship between proposal features and instance features in cosine space. In addition, most existing algorithms, although based on a dual branch architecture, only utilize one memory bank for storing instance features, ignoring a large amount of background noise in inaccurate proposal features, which affects the quality of subsequent pseudo label generation.

To solve the above problems, this paper proposes a novel method for optimizing label assignments in weakly supervised person search. In the clustering phase, we propose *Context Aware Clustering* module, which introduces new constraints to the DBSCAN clustering algorithm by constructing a graph of multiple pedestrians within each scene. The graph's matching results, derived from bipartite graph matching, replace the traditional threshold-based decision process. During the training phase, we propose *Prototype Matching* module, which first crop given bounding boxes to obtain pedestrian instance information. We then use an encoder to generate corresponding feature encodings as instance features, which are subsequently used to calculate the IoU distance and Cosine distance with proposals generated by RPN. By summing these two distances, we formulate a cost matrix based on optimal transport theory, yielding label assignments from a global perspective. Finally, considering that proposal features with background noise are also updated in the memory bank, which could affect subsequent label assignments, we innovatively propose a *Dual Memory Bank Enhancement* module that separates instance features from proposal features, significantly enhancing the accuracy of label assignments. Our contributions are summarized as follows:

- We redefine the clustering strategy for person search by incorporating graph-based constraints into DBSCAN. This approach leverages the relationships between multiple pedestrians in a scene, utilizing bipartite graph matching to replace the conventional threshold-based selection.
- We employ optimal transport theory to calculate label assignments. This method integrates the IOU and cosine distances into a unified cost matrix, providing a comprehensive and robust mechanism for matching instance features with proposals.
- We propose a Dual Memory Bank Enhancement module, and thus address the challenge of background noise in proposal features influencing the accuracy of label assignments.

Related Work

Weakly Supervised Person Search Weakly supervised person search extends the person ReID task by locating individuals in scenes using only bounding box annotations, thereby reducing labeling efforts and enhancing scalability. Traditional person search methods are categorized into two-step (Wang et al. 2020; Dong et al. 2020) and one-step (Yu et al. 2022; Cao et al. 2022; Han, Ko, and Sim 2021; Yan et al. 2023) approaches. Two-step methods sequentially employ separate networks for person detection and ReID feature extraction, as exemplified by DPM (Zheng et al. 2017), which introduced a large-scale benchmark and a confidence-weighted similarity method to suppress false positives. One-step approaches integrate detection and ReID within a unified architecture, improving efficiency and processing speed (Yan et al. 2023; Cao et al. 2022).

In the realm of weakly supervised person search, several frameworks leverage Siamese networks and contrastive learning to compensate for the lack of identity labels. CGPS (Yan et al. 2022) integrates detection with memory and scene-level context to enhance discriminative capabilities and clustering accuracy. R-SiamNets (Han et al. 2021a) employs instance-level consistency and cluster-level contrastive losses to maintain feature uniformity and discriminativeness. DICL (Wang et al. 2024) addresses spatial and occlusion variances through specialized contrast modules, while SSL (Wang et al. 2023) handles scale variations with self-similarity-driven learning and dynamic multi-label prediction. These methods effectively improve feature representation and identity matching through sophisticated contrastive strategies. Our proposed approach builds on the Siamese network architecture, focusing on optimizing label generation and update mechanisms through dynamic multi-label prediction and refined label optimization, thereby further enhancing detection and re-identification performance in weakly supervised settings.

Unsupervised Person ReID The main core issue of weakly supervised person search is the lack of pedestrian labels, so feature learning tasks can be abstracted as unsupervised ReID tasks, which refers to identifying individuals across different scenes or camera views without relying on labeled training data. These approaches are generally classified into two categories. The first is the generative network based methods, which leverage Generative Adversarial Networks (GANs) to adaptively learn domain-invariant features across different visual domains, either by transforming the visual style of images or by separating identity-related and unrelated features in the feature space. Another approach employs pseudo labels, which apply Expectation Maximization (EM) to refine pseudo labels generated by clustering algorithms and optimize network parameters with pseudo labels iteratively. Most recent works focus on pseudo-label generation and framework design. Specifically, SPCL(Ge et al. 2020) treats each cluster and outlier as a single class while performing contrastive learning based on a hybrid memory containing cross-dataset features. More recent contributions (Dai et al. 2022; Zhang et al. 2022; Yin et al. 2023) maintain the use of contrastive learning but introduce novel

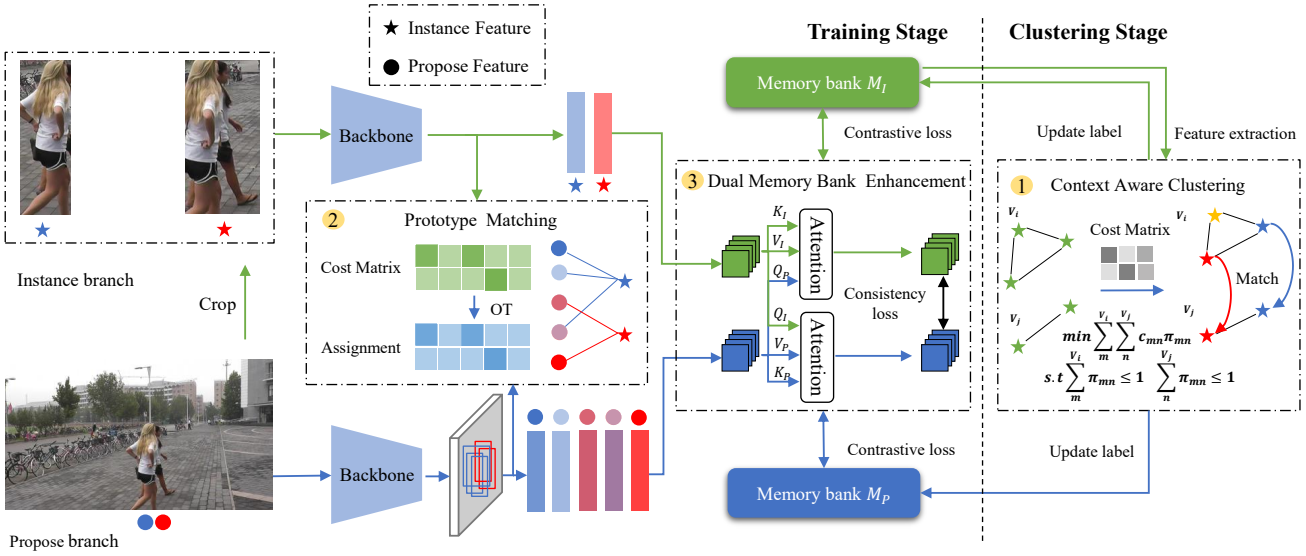


Figure 2: The overall of our Optimizing Label Assignment (OLA). It consists of the main branch, instance branch, and dual memory bank. The main branch processes scene images to detect persons and extract their features, subsequently updating the memory bank M_P with these features. The instance branch uses a person instance image cropped from the scene image according to the bounding box label. The instance branch takes it as input and updates it to the memory bank M_I. Before the start of each epoch, the model uses Context Aware Clustering leveraging features in M_I to generate pseudo labels for the training process.

considerations, such as identity centroids for more precise camera-based clustering, enhancing cluster consistency, and improving the reliability of pseudo label generation. In this paper, because most prototypes only have a single instance, we choose the instance updating method based SPCL(Ge et al. 2020) as the feature learning method.

Proposed Method

Preliminaries

An overview of our training framework is shown in Fig. 2, given training set $X = \{x_1, x_2, ..., x_T\}$ of T images, weakly supervised person search aims to find the network parameter θ to obtain the best embedding representation $Z = \{z_1, z_2, ..., z_q\}$ of the person with bounding box annotations $B = \{b_1, b_2, ..., b_q\}$. We firstly use the *Context Aware Clustering* algorithm to generate pseudo labels $Y = \{y_1, y_2, ..., y_q\}$. During the training phase, we assign pseudo labels to proposals by introducing *Prototype Matching* and finally update embedding representation through *Dual Memory Bank Enhancement*.

Context Aware Clustering

We present a novel Context Aware Clustering (CAC) algorithm in the cluster phase for integrating context information into the matching process. Traditional clustering algorithms in weakly supervised person search tasks are typically instance-level strategies. These algorithms form clusters from embeddings that exceed a certain threshold

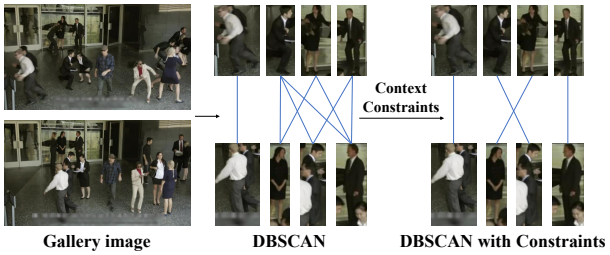


Figure 3: Cluster result w/o Context Aware Clustering.

and merge clusters that meet specified criteria. This approach aligns with the standard DBSCAN algorithm, where density-based spatial clustering forms the core methodology. However, as shown in Fig. 3, it may fail when there are multiple people with very similar appearances in the same scene image. This is because the strategy does not accommodate context-specific constraints, such as ensuring that identical pedestrians do not appear simultaneously in the same scene image. We replace the threshold based matching method with the bipartite graph based matching method. In this way, there will never be two similar pedestrians in a scene image.

For any image x_i, we define the V_i as the embedding set of all people in x_i. Randomly selecting two sets V_i, V_j, we firstly build a complete bipartite graph $G = (\mathbb{V}, \mathbb{E})$, in which $\mathbb{V} = V_i \cup V_j$. The graph has the following properties:

Algorithm 1: Context-Aware Clustering (CAC)

Require: X: images with detected instances, Z: feature vectors, β: adjustment factor

for each image $x \in X$ **do**
  for each pair of distinct instances (z_i, z_j) within x **do**
    $sim(z_i, z_j) \leftarrow 0$ ▷ Set intra-image similarity to zero
  end for
end for
$C \leftarrow$ DBSCAN(Z, sim)
for each cluster C_i in C **do**
  if C_i contains multiple instances from the same image **then**
    for each pair of images (V_i, V_j) in C_i **do**
      Construct bipartite graph G with instances from V_i and V_j
      $M \leftarrow$ K-Matching(G)
      for each pair (v_i, v_j) in G **do**
        if $(v_i, v_j) \in M$ **then**
          $sim(v_i, v_j) \leftarrow sim(v_i, v_j) + \beta$
        else
          $sim(v_i, v_j) \leftarrow sim(v_i, v_j) - \beta$
        end if
      end for
    end for
  end if
end for
$C' \leftarrow$ DBSCAN(Z, sim)
return C'

- For every two vertices $v_1 \in V_i$ and $v_2 \in V_j$, if the cosine distance between the vertices of v_1 and v_2 below 0.5, an edge is established in $\mathbb{E}$.
- No edge has both endpoints in the same set of vertices.
- Each edge $e \in \mathbb{E}$ is weighted by the cosine distance between the corresponding vertices, representing the cost c required to match them.

We establish the optimal matching result by applying the Kuhn-Munkres (K-M) algorithm (Kuhn 1955) to minimize the weighted sum of edge costs:

$$\min \sum_{m}^{V_i} \sum_{n}^{V_j} c_{mn} \pi_{mn}, \tag{1}$$

due to contextual constraints in the two images:

$$\text{s.t.} \quad \sum_{m}^{V_i} \pi_{mn} \leq 1, \sum_{n}^{V_j} \pi_{mn} \leq 1, \tag{2}$$

where, c_{mn} denotes the cosine distance between vertices v_m and v_n, representing the cost of matching them, while the binary variable π_{mn} indicates whether a match is established between these vertices. The constraints ensure that each vertex in V_i is matched to at most one vertex in V_j and vice versa, thereby enforcing a unique one-to-one correspondence and allowing the Kuhn-Munkres algorithm to minimize the total matching cost effectively.

First, for all instances within the same image, we set their pairwise similarity to zero to prevent intra-image instances from influencing the clustering process. We then apply a clustering algorithm (e.g., DBSCAN) on the instance-level Re-ID features to obtain initial clusters. Next, we identify problematic clusters, which are clusters where multiple instances from the same image are assigned to the same cluster center—indicating erroneous groupings. For each problematic cluster, we examine pairs of images within the cluster and construct a complete bipartite graph between their instances. Using the Kuhn-Munkres (K-M) algorithm, we find the optimal maximum-weight matching in this graph. Based on the matching results, we adjust the similarities by increasing the similarity by β for matched pairs and decreasing it by β for unmatched pairs. The proposed CAC algorithm is described in Algorithm. 1. Although we expanded the computational parameters, the time complexity does not increase significantly when the number of clusters is large.

Prototype Matching

In the context of person search, supposing there are m targets and n proposal generated by the RPN network for an input image x_t, We need to assign the most likely label to each proposal for subsequent network learning. Inspired by OTA(Ge et al. 2021), we task this problem as Optimal Transport and view each gt as a supplier s_i who holds k units of positive labels ($i.e., s_i = k, i = 1, 2, ..., m$), and each anchor as a demander d_j who needs one unit of the label ($i.e., d_j = 1, j = 1, 2, ..., n$). We define the cost c_{ij} for transporting one unit of the positive label from gt_i to proposal a_j as the weighted summation of IoU distance and cosine distance:

$$c_{ij}^{fg} = L_{reg}(P_j^{box}(\theta), G_i^{box}) + \alpha L_{id}(P_j^{id}(\theta), G_i^{id}), \tag{3}$$

where θ stands for the model's parameters. P_j^{id} and P_j^{box} denote predicted ID embedding and bounding box for a_j. G_i^{id} and G_i^{box} represent the ground truth instance embedding, cropped according to its bounding box, and the bounding box itself, for the ground truth instance gt_i. L_{reg} and L_{id} represent the IoU Loss and SPCL Loss (Ge et al. 2020). α is the balanced coefficient.

Besides positive assigning, a large set of anchors are treated as negative samples during training. Because the optimal transportation need involves all anchors, we introduce another supplier, background, which only provides negative labels. In a standard OT problem, the total supply must equal the total demand. We thus set the number of negative labels that the background can supply as $n - m \times k$. We randomly crop b proposals in the background of the scene image and generate embeddings through instance branching, and aggregate them to represent a background embedding G_{bg}^{id}. The cost for transporting one unit of negative label from background to a_j is defined as:

$$c_j^{bg} = \sum_{i}^{k} IoU(P_j^{box}(\theta), G_i^{box}) + \alpha L_{id}(P_j^{id}(\theta), G_{bg}^{id}), \tag{4}$$

where the first term means the sum of IoU between proposal a_j and gt_i. If it is background, the first term is calculated as

zero. Concatenating this $c^{bg} \in \mathbb{R}^{1\times n}$ to the last row of $c^{fg} \in \mathbb{R}^{m\times n}$, we can get the complete form of the cost matrix $c \in \mathbb{R}^{m+1\times n}$. The supplying vectors should be correspondingly updated as:

$$s_i = \begin{cases} k, & \text{if } i \leq m \\ n - m \times k, & \text{if } i = m+1 \end{cases}. \tag{5}$$

As we already have the cost matrix c, supplying vector $s \in \mathbb{R}^{m+1}$ and demanding vector $d \in \mathbb{R}^{n}$, the optimal transportation plan $\pi^* \in \mathbb{R}^{m+1\times n}$ can be obtained by solving this OT problem via the off-the-shelf Sinkhorn-Knopp Iteration (Cuturi 2013). After getting π^*, one can decode the corresponding label assigning solution by assigning each anchor to the supplier who transports the largest amount of labels to them.

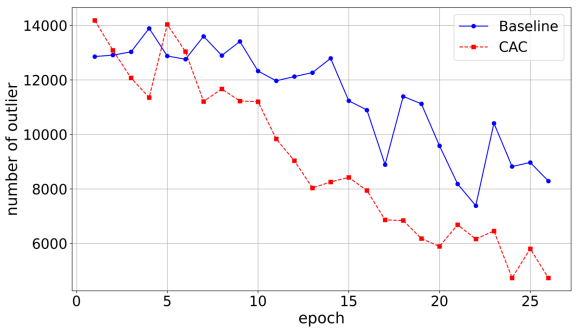
Intuitively, the appropriate number k of positive anchors for each gt should be different and based on many factors like objects' sizes, scales, occlusion conditions, etc. As it is hard to directly model a mapping function from these factors to the positive anchor's number, we propose a simple but effective method to roughly estimate the appropriate number of positive anchors for each gt based on the cost values between predicted and ground truth boxes. Specifically, for each gt, we select the top q predictions according to cost values. These cost values are summed up to represent this gt estimated number of positive anchors.

Dual Memory Bank Enhancement

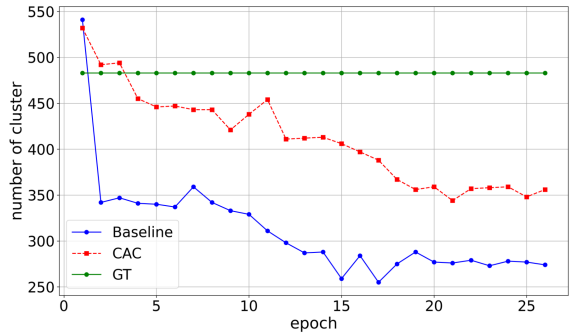
Intuitively, the prototype stored in the memory bank is an essential latent variable for weakly supervised pedestrian search training, and its update method affects the network's learning process. The traditional single memory bank effectively utilizes contextual information by storing the average features from the two branches as the updated amount. However, this process introduces background noise due to inaccurate proposals from the RPN network. Moreover, computing the contrastive loss becomes challenging with a limited number of pedestrian instance images. We propose a simple but effective method to separate the update process of instance features and proposal features. As shown in Fig. 2, we first add an additional memory bank to the instance branch and perform cross-attention between the instance and proposal branch to allow information communication between the instance branch and the proposal branch. The cross attention is defined as:

$$\begin{aligned} \text{feat}_P &= \text{Softmax}\left(\frac{Q_I K_P^T}{\sqrt{d}}\right) V_P, \\ \text{feat}_I &= \text{Softmax}\left(\frac{Q_P K_I^T}{\sqrt{d}}\right) V_I, \end{aligned} \tag{6}$$

where Q_I and Q_P represent query embedding in the instance branch and proposal branch, respectively. K_I and K_P represent key embedding, V_I and V_P represent value embedding, and d represents dimension scaling factor. The role of cross attention is to effectively exchange information between instance branches and proposal branches by focusing on the features of the other branch, which helps reduce feature pollution caused by background noise. When predicting pseudo labels, we utilize the memory banks of instance branches for prediction and ultimately allocate the labels to two memory banks.



(a) Outlier numbers with different epochs.



(b) Cluster numbers with different epochs.

Figure 4: Outlier and Cluster numbers across different epochs.

Experiments

Datasets and Settings

CUHK-SYSU CUHK-SYSU dataset (Xiao et al. 2017) is a large-scale person search dataset, which is composed of urban scene pictures and movie snapshots. There are 18, 184 images with 96, 143 annotated bounding boxes, including 8, 432 labeled identities. The training set consists of 11, 206 images with 5, 532 identities and several unlabeled ones. The testing set has 6, 978 gallery images and 2, 900 probe images.

PRW PRW dataset (Zheng et al. 2017) is captured by six spatially disjoint cameras in the university. It consists of 11, 816 frames with 43, 110 annotated bounding boxes, among which 34, 304 are assigned with 932 identity labels, and the rest are unlabeled. The training set contains 5, 704frames with 482 identities, and the testing set includes 6, 112 gallery images and 2,057 queries with 450 identities.

Evaluation Protocols Our experiments employ the standard evaluation metrics in person search. One is the cumulative matching cure (CMC), which is inherited from the person ReID. A candidate is counted if the IoU with ground truth is greater than 50%. The other is the mean Average Precision (mAP), which is inspired by the object detection

PM	CAC	DMBE	ReID		Detection	
			mAP	Top-1	Recall	mAP
			35.4	81.1	62.1	55.3
✓			36.4	81.5	64.6	56.2
✓	✓		37.2	81.7	65.0	56.7
✓	✓	✓	**38.1**	**82.0**	**65.1**	**57.1**

Table 1: Ablation study over the PRW dataset by incrementally adding our novel contributions to the baseline.

task. We compute an averaged precision (AP) for each query based on the precision and recall curve. Then, the mAP is calculated by averaging the APs across all the queries.

Implementation Details

We employ Faster-RCNN (Ren et al. 2015) released by OpenMMLab (Chen et al. 2019b) as our backbone network, containing the ImageNet (Krizhevsky, Sutskever, and Hinton 2012) pretrained ResNet-50 (He et al. 2016). The main path takes the scene images as inputs, jointly training the detection and recognition. The RoIalign is applied on the backbone with the proposals RoIs to obtain the person features, followed by a fully connected (FC) layer after flattening. The cropped and resized images are taken as inputs for the instance path. FC layer is directly applied to the backbone to obtain the person's features.

The scene images are resized to 1500×900, and cropped images are rescaled to 224×96. The batched Stochastic Gradient Descent (SGD) optimizer is used with a momentum of 0.9. The weight decay factor for L2 regularization is set to 5×10^{-4}. We use a mini-batch size of 4 for the main batch and a batch size of 16 for asynchronous data. The initial learning rate is 1×10^{-3}. We set the adjustment factor β to 0.2 and the label assignment parameter k to 3. The model is trained for 26 epochs with the learning rate multiplied by 0.1 at 16 and 22 epochs. All experiments is implemented on the PyTorch framework, and the network is trained on the NVIDIA RTX 4090.

Ablation Study

Effectiveness of Each Component We evaluate the effectiveness of our Optimizing Label Assignment (OLA) method in Tab. 1, where PM stands for Prototype Matching, CAC for Context-Aware Clustering, and DMBE for Dual Memory Bank Enhancement. The baseline model achieves 35.4% mAP and 81.1% Top-1 accuracy on PRW. Incorporating PM improves mAP by 1.0% and Top-1 by 0.4%, along with significant detection enhancements, increasing recall and mAP by 1.5% and 0.9%, respectively. Adding CAC further boosts the baseline by 0.8% mAP and 0.2% Top-1, highlighting the value of context information in weakly supervised person search. Integrating DMBE results in an additional 0.9% mAP and 0.3% Top-1 improvement. Overall, OLA achieves a total improvement of 2.7% in mAP and 0.9% in Top-1 accuracy, demonstrating the effectiveness of our asynchronous learning approach in weakly supervised person search.

Method	CUHK-SYSU		PRW	
	mAP	Top-1	mAP	Top-1
Two-Step				
DPM(Zheng et al. 2017)	-	-	20.5	48.3
MGTS(Chen et al. 2018)	83.0	83.7	32.6	72.1
RDLR(Han et al. 2019)	93.0	94.2	42.9	70.2
IGPN(Dong et al. 2020)	90.3	91.4	47.2	87.0
TCTS(Wang et al. 2020)	93.9	95.1	46.8	87.5
One-step				
OIM (Xiao et al. 2017)	75.5	78.7	21.3	49.4
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
NAE (Chen et al. 2020b)	91.5	92.4	43.3	80.9
DMRNet (Han et al. 2021b)	93.2	94.2	46.9	83.3
SeqNet (Li and Miao 2021)	93.8	94.6	46.7	83.4
DKD (Zhang et al. 2021)	93.1	94.2	50.5	87.1
OIMNet++(Lee et al. 2022)	93.1	93.9	46.8	83.9
PGS(Kim et al. 2021)	92.3	94.7	44.2	85.2
AlignPS(Yan et al. 2021)	93.1	93.4	45.9	81.9
PSTR(Cao et al. 2022)	93.5	95.0	49.5	87.8
COAT(Yu et al. 2022)	94.2	94.7	53.3	87.4
weakly person search				
CGPS(Yan et al. 2022)	80.0	82.3	16.2	68.0
R-SiamNet (Han et al. 2021a)	86.0	87.1	21.2	73.4
SSL (Wang et al. 2023)	87.4	88.5	30.7	80.6
DICL (Wang et al. 2024)	87.4	88.8	35.5	80.9
DICL* (Wang et al. 2024)	86.8	88.1	35.4	81.1
Ours(OLA)	**87.8**	**89.3**	**38.1**	**82.0**

Table 2: Comparison with state-of-the-arts regarding mAP and Top-1 accuracy on CUHK-SYSU and PRW test sets. All methods are implemented with the same backbone ResNet50. * represents our reproduced results.

Effectiveness of Context Aware Clustering. We count the number of outliers at the end of each epoch, as shown in Fig. 4a. Leveraging the CAC module reduces the number of outliers, which enhances the model's focus on relevant features and improves learning efficiency. Furthermore, Fig. 4b demonstrates that the CAC (represented by the red line) achieves a cluster count that more accurately approximates the actual label distribution (green line), indicating enhanced model accuracy. The outcomes of our experiments conclusively demonstrate that our proposed Clustering algorithm significantly improves model performance, streamlines optimization processes, and accelerates iteration speeds, thereby underscoring its importance in weakly supervised person search tasks.

Comparison with the State-of-the-Arts

Qualitative Results Fig. 5 demonstrates the effectiveness of our Optimizing Label Assignment algorithm on the challenging CUHK-SYSU test sets. Our method accurately detects and recognizes individuals across diverse gallery images, adapting to variations in poses, scales, backgrounds, and visibility. In the first row, the algorithm successfully identifies a pedestrian in a highly obstructed environment,

Figure 5: Qualitative results of our method on CUHK-SYSU test set, where the red box represents the query and the green box represents the search result in the gallery image. Our method matches the query persons in different scenes.

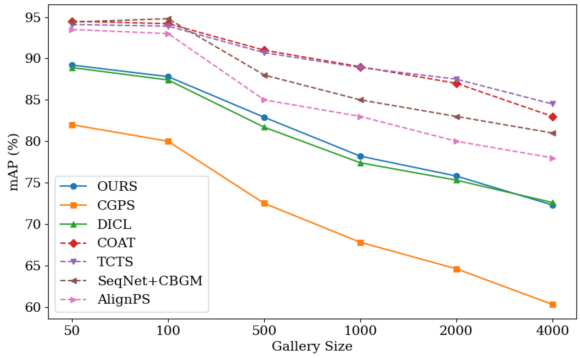
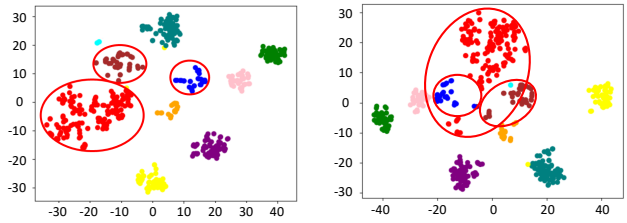


Figure 6: Performance with different gallery images.

showcasing its ability to handle complex spatial contexts. The last row highlights robust recognition despite significant scale changes, while the middle rows illustrate consistent identification under adverse lighting conditions. These qualitative results affirm the robustness and practical effectiveness of our approach in real-world scenarios.

Results on CUHK-SYSU Tab. 2 shows that our method achieves the highest mAP of 87.8% and Top-1 accuracy of 89.3% on CUHK-SYSU with a gallery size of 100, surpassing all existing weakly supervised person search methods. To verify robustness, we tested gallery sizes from 50 to 4,000. Fig. 6 compares mAP, indicating that our method consistently outperforms other weakly supervised approaches as gallery size increases.

Results on PRW As shown in Tab. 2, our method achieves 38.2% mAP and 83.3% Top-1 accuracy, outperforming all existing weakly supervised methods by a significant margin. Additionally, we use t-SNE to visualize feature distributions on a subset of the PRW dataset with 10 classes and 511 pedestrians, where different colors represent distinct classes. Fig. 7(a) displays the baseline model's features, showing large intra-class and small inter-class distances. In contrast, Fig. 7(b) illustrates that our method achieves more cohesive



(a) baseline (b) ours(OLA)

Figure 7: t-SNE feature visualization on the part of the PRW training set. Colors denote different personal identities.

feature aggregation within each category, demonstrating the effectiveness of our Optimizing Label Assignment (OLA).

Conclusion

We proposed Optimizing Label Assignment (OLA) for weakly supervised person search, which incorporates graph-based constraints into the DBSCAN clustering process and utilizes optimal transport theory for robust label assignment during training, representing a fundamental shift towards more accurate and efficient person search systems. The introduction of a decoupled memory bank further refines the label assignment process by minimizing the influence of background noise.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62372348 and 62036007, in part by the Key Research and Development Program of Shaanxi under Grant 2024GX-ZDCYL-02-10, in part by Shaanxi Outstanding Youth Science Fund Project under Grant 2023-JC-JQ-53, in part by the Shaanxi Province Core Technology Research and Development Project under grant 2024QY2-GJHX-11, in part by the Fundamental Research Funds for the Central Universities under Grant QTZX24080 and QTZX23042.

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*, 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 Conference on Computer Vision*, 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.

Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W.; et al. 2019a. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4974–4983.

Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. 2019b. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26.

Dai, Z.; Wang, G.; Yuan, W.; Zhu, S.; and Tan, P. 2022. Cluster contrast for unsupervised person re-identification. In *Proceedings of the Asian Conference on Computer Vision*, 1142–1160.

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

Ge, Y.; Zhu, F.; Chen, D.; Zhao, R.; et al. 2020. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Advances in Neural Information Processing Systems*, 33: 11309–11321.

Ge, Z.; Liu, S.; Li, Z.; Yoshie, O.; and Sun, J. 2021. Ota: Optimal transport assignment for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 303–312.

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.; Su, K.; Yu, D.; Yuan, Z.; Gao, C.; Sang, N.; Yang, Y.; and Wang, C. 2021a. Weakly supervised person search with region siamese networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12006–12015.

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.; Gao, C.; Sang, N.; and Yang, Y. 2021b. Decoupled and memory-reinforced networks: Towards effective feature learning for one-step person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1505–1512.

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

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15013–15022.

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*, 4865–4874.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2): 83–97.

Lee, S.; Oh, Y.; Baek, D.; Lee, J.; and Ham, B. 2022. Oimnet++: Prototypical normalization and localization-aware learning for person search. In *Proceedings of the European Conference on Computer Vision*, 621–637.

Li, Y.; He, J.; Zhang, T.; Liu, X.; Zhang, Y.; and Wu, F. 2021. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2898–2907.

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.

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. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*.

Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C.; et al. 2021. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14454–14463.

Wang, B.; Yang, Y.; Wu, J.; Qi, G.-j.; and Lei, Z. 2023. Self-similarity Driven Scale-invariant Learning for Weakly Supervised Person Search. *arXiv:2302.12986*.

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/CVF Conference on Computer Vision and Pattern Recognition*, 3415–3424.

Yan, Y.; Li, J.; Liao, S.; Qin, J.; Ni, B.; Lu, K.; and Yang, X. 2022. Exploring visual context for weakly supervised person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3027–3035.

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.

Yan, Y.; Li, J.; Qin, J.; Zheng, P.; Liao, S.; and Yang, X. 2023. Efficient person search: An anchor-free approach. *International Journal of Computer Vision*, 1–20.

Yin, J.; Zhang, X.; Ma, Z.; Guo, J.; and Liu, Y. 2023. A real-time memory updating strategy for unsupervised person re-identification. *IEEE Transactions on Image Processing*.

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.

Zhang, X.; Li, D.; Wang, Z.; Wang, J.; Ding, E.; Shi, J. Q.; Zhang, Z.; and Wang, J. 2022. Implicit sample extension for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7369–7378.

Zhang, X.; Wang, X.; Bian, J.-W.; Shen, C.; and You, M. 2021. Diverse knowledge distillation for end-to-end person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3412–3420.

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.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *International Conference on Learning Representations*.