Learning Scene-Pedestrian Graph for End-to-End Person Search

Zifan Song, Cairong Zhao, Guosheng Hu, *Senior Member, IEEE*, and Duoqian Miao

***Abstract*—Person search aims to find specific persons from visual scenes, including two subtasks, pedestrian detection, and person reidentification. The dominant fashion in this area is end-to-end networks that focus on analyzing the foreground (i.e., pedestrian) while ignoring the background (i.e., scene) information. However, the scene information often offers useful clues for person search. For example, pedestrians normally appear on the road rather than the top of a tree, and pedestrians appearing at the same location are likely to have similar occlusions. The interplay between the pedestrians and scenes can potentially improve the performance. In this article, a novel scene-pedestrian graph (SPG) is proposed, which can explicitly model the interplay between the pedestrians and scenes. To polish the quality of pedestrian bounding boxes, we pioneer a strategy of using the high-quality pedestrian bounding box to guide the low-quality one in the same scene. In addition, we design a contextual and temporal graph matching algorithm to effectively utilize the contextual and temporal information present in the constructed SPG to improve the performance of pedestrian matching. Benefiting from the robustness on complex scenes, our model achieves promising performance over the state-of-the-art methods on two popular person search benchmarks, CUHK-SYSU and PRW.**

***Index Terms*—Deep learning, graph neural networks, identification of persons, machine vision.**

Manuscript received 30 August 2022; revised 27 January 2023 and 22 May 2023; accepted 13 July 2023. Date of publication 9 August 2023; date of current version 19 January 2024. This work was supported in part by the National Natural Science Fund of China under Grant 62076184, Grant 61976158, Grant 61976160, Grant 62076182, and Grant 62276190; in part by Fundamental Research Funds for the Central Universities and State Key Laboratory of Integrated Services Networks (Xidian University); in part by the Shanghai Innovation Action Project of Science and Technology under Grant 20511100700; and in part by the Shanghai Natural Science Foundation under Grant 22ZR1466700. Paper no. TII-22-3696. *(Corresponding author: Cairong Zhao.)*

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval. We evaluate our proposed approach on two widely used public benchmark datasets for person search, CUHK-SYSU and PRW. And the human subjects are involved in these two datasets.

Zifan Song, Cairong Zhao, and Duoqian Miao are with the Department of Computer Science and Technology, Tongji University, Shanghai 200092, China (e-mail: 2111139@tongji.edu.cn; zhaocairong@tongji.edu.cn; dqmiao@tongji.edu.cn).

Guosheng Hu is with the Oosto, BT3 9DT Belfast, U.K. (e-mail: huguosheng100@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TII.2023.3298473.

Digital Object Identifier 10.1109/TII.2023.3298473

I. Introduction

PERSON search [1] aims at locating specific pedestrians in scene images captured by different cameras, consisting of two primary subtasks: 1) pedestrian detection [2]; and 2) pedestrian reidentification (re-ID) [3], [4], [5], [6], [7]. Pedestrian detection detects all the pedestrians in scene images and generates their bounding boxes (BBoxes). Pedestrian re-ID is an identity-matching task that assumes pedestrian detection has been accurately performed. As a combination of these two tasks, person search is more relevant to real-world applications. As a result, various real-world factors pose significant challenges, including different viewpoints cross cameras, occlusions, etc.

Person search is oriented to open outdoor scenes, and existing works focus on foreground (pedestrian) features in scene images, often disregarding background (scene) information. In the person search task, various cameras are stationary and capture scene frames from fixed positions and directions. Therefore, video frames captured under the same camera tend to have similar scene information. As illustrated in Fig. 1(a), the four images are obtained from the same camera. Despite the varying pedestrians, the background (e.g., bikes and buildings) remains constant. To model the specific confidence distribution of pedestrian BBoxes in the scene, we highlight the pedestrian (dynamic) and background (still) area in blue and white colors, respectively. On the other hand, in Fig. 1(b), existing BBox correction methods [8], [9] sometimes fail to perform accurate corrections in complex scenes. In the case on the right side of Fig. 1(b), the model generates an inappropriate correction owing to the similarity between the features of the pedestrian's legs and the background, whereas the case in the left side is corrected properly. Since the scenes are identical, we are inspired to improve the BBox correction by enabling the failed case (red BBox) to learn from the successful case (green BBox) in the same scene with similar positions.

Motivated by the above observations and analysis, we propose a novel scene-pedestrian graph (SPG), to explicitly model the interplay between pedestrians and scenes. Specifically, we define two types of nodes in our SPG: 1) pedestrian nodes; and 2) scene nodes. A pedestrian node contains detailed information modeling a person, including the BBox features, IoU with the target pedestrian, identity ID, latest BBox correction, and temporal information (e.g., video clip number and frame number). To model scene nodes, we divide the scene image into blocks and treat each block as a node. Then, we encode the position information of each block and the corresponding pedestrian BBox confidence into the node. In addition, we provide adaptive scene image block

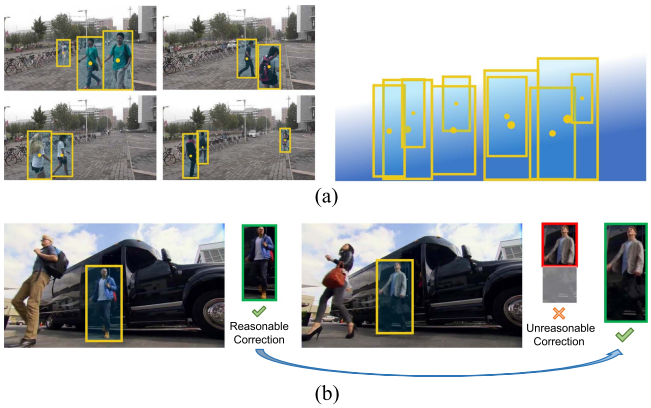

Fig. 1. Analysis of scene information. (a) Pedestrian BBoxes in the same scene usually appear in the similar area, e.g., roads highlighted by the blue color. (b) Pedestrian BBoxes corrections at the similar locations can learn from the reasonable ones, e.g., the correction in the green BBox can guide the red one.

sizes for different scenes, which can learn different block sizes to adjust to the interactions between pedestrians and scenes. Our SPG can solve two problems based on the learned interplay: 1) removing the wrong BBoxs that appear in unreasonable areas [e.g., the white area in Fig. 1(a)]; and 2) performing accurate BBox corrections by using the high-quality BBoxes to guide the low-quality ones in the same scene [Fig. 1(b)].

Accurate pedestrian detection, achieved by SPG or other methods, does not always result in accurate pedestrian re-ID in complex scenes due to cross-camera viewpoint differences and occlusions. Relying solely on the similarity between the query pedestrian and the candidate pedestrian is not reliable. In addition, the person search datasets have low homogeneity (i.e., many pedestrians identities but few samples for each identity). This makes it challenging to learn a robust pedestrian feature representation. Thus, we propose to use the encoded information (e.g., contextual and temporal information) in our SPG and design a contextual and temporal graph matching (CTGM) algorithm to enhance the performance of re-ID. Our matching strategy not only considers the key pedestrian but also the surrounding people. Moreover, we propose to incorporate temporal information into the matching process. Specifically, given a query, since pedestrians typically move slowly, we reduce the matching scores of candidates who are temporally adjacent but spatially distant from the query but temporally adjacent. As a result, our method can reduce the re-ID misclassification caused by multiple people with similar appearance.

Our main contributions are summarized as follows.

1) We propose a new SPG to model the interactions between pedestrians and scenes to achieve high-quality pedestrian detection.
2) For pedestrian matching of re-ID, we propose a CTGM algorithm which models both the contextual and temporal information extracted from SPG, achieving robust matching for walking pedestrians in complex scenes.
3) Extensive experimental results on two popular person search datasets CUHK-SYSU and PRW demonstrate that our method performs favorably against the state-of-the-art methods. On the PRW dataset, in particular, our method significantly outperforms the state-of-the-art ones by 2.24% in terms of the rank-1 score.

II. Related Work

A. Person Search

Person search aims to locate the query person in scene images, consisting of pedestrian detection and person re-ID. Existing works can be mainly classified into two categories: 1) two-stage approaches and 2) end-to-end approaches. Two-stage approaches train two independent models for pedestrian detection and person re-ID, while end-to-end approaches combine the two subtasks in an unified model.

For two-stage methods, Zhao et al. propose to refine the pedestrian matching scores with the detection confidence and construct a dataset for person search. RDLR [8] introduced a differentiable ROI layer to combine the detection head and the re-ID head, trained with the proxy triplet loss. Yao et al. [10] designed the OR similarity to utilize the objectness and repulsion information.

For end-to-end methods, the earliest approach dates back to 2017 [11]. The original end-to-end person search algorithm implements a single convolutional neural network including a pedestrian proposal net and an identification net, using online instance matching (OIM) loss function to train the network. Inspired by this work, many improvements on its basis are subsequently derived. OIAM [12] introduced two independent CNN blocks: the first one connects to the nonlocal layer to generate image features and the second one sends feature vectors to the fully connected layer to generate candidate proposals for classification and regression. Bharti Munjal et al. [13] proposed the idea of QSSE-Net, QRPN, and QSim-Net three-seeded networks to implement query-adaptive search. Chen et al. [14] proposed the hierarchical online instance matching (HOIM) loss, aiming to explicitly integrate the hierarchy of pedestrian detection and pedestrian re-ID into OIM loss to better classify pedestrians and backgrounds. Chen et al. [15] introduced the norm-aware embedding (NAE) method and the extended pixel-based method NAE+ to address the misalignment problem. For the occlusion problem in complex scenes, Zhong et al. designed an APNet to extract and align distinguished part features, which naturally formulates the re-ID as a partial feature matching process [9]. Li et al. [16] noticed the performance of previous end-to-end framework is limited by inferior features and proposed a sequential end-to-end network (SeqNet) to solve the pedestrian detection and re-ID in turn. Yan et al. [17] proposed the first anchor-free model AlignPS to simplify the framework for person search, where pedestrian detection and re-ID are jointly addressed by a one-step model.

B. GCN for Visual Tasks

Different from CNN, graph convolution network (GCN) [18] specially deals with irregular graph data and achieves the process of filtering in the frequency domain. The earliest prominent

Fig. 2. Illustration of the person search task. The blue and yellow bounding boxes denote the query person and the gallery persons, respectively. The green and red bounding boxes denote the correct and false matches, respectively.

study of GCN is based on graph spectral theory [19] and its key idea is node feature aggregation, i.e., updating the node/edge features based on their relationships. In recent years, GCN is used in a wide range of visual works [20], [21] and has proven to be effective. For pedestrian re-ID, Chen et al. [22] proposed heterogeneous graph embeddings to preserve more abundant cross-modal information, showing strong performance and superiority on cross-modality person re-ID. For person search, previous work [23] constructs graphs modeling the interplay between probe-gallery pairs to improve the robustness of matching, but ignoring the scene and temporal information. In this work, we propose an SPG to model the interplay between the pedestrian BBoxes and scene information, utilizing both the contextual and temporal information.

III. Proposed Method

In this section, we first describe the problem formulation of person search and present our framework of person search. Then, we introduce our proposed SPG and CTGM algorithm in detail.

A. Problem Formulation

As illustrated in Fig. 2, given a query pedestrian q in the query image Q and a set of gallery images $S = \{G_1, G_2, \ldots, G_N\}$, the purpose of person search is to detect a collection of pedestrian BBoxes B in S and then find the best matching pedestrian of q in B as the output.

Pedestrian detection is expected to generate suitable BBoxes to serve the re-ID; however, most existing works generate low-quality BBoxes when facing issues, such as occlusion. Our solution is to construct an SPG with a novel node representation and model the interplay between the pedestrian BBoxes and scene information. We eliminate pedestrian BBoxes in unreasonable locations by assigning the confidence distribution of BBoxes for each scene. In addition, we use high-quality pedestrian BBox corrections to guide low-quality ones.

B. End-to-End Framework for Person Search

We adopt SeqNet [16] as our baseline for and the overview of our approach is shown in Fig. 3. The overall architecture includes a standard faster region-based convolutional neural network (R-CNN) [24] head to generate pedestrian BBoxes and a second head to further fine-tune the BBoxes, solving pedestrian detection and re-ID, respectively. Between these two heads, we construct our SPG by extracting pedestrian/scene information from the first head/scene images, respectively. After node feature aggregation, we select the relevant subgraphs from our constructed SPG and feed their features into both the proposed CTGM algorithm and the second head to complete the re-ID task.

C. Scene-Pedestrian Graph

Pedestrian detection is expected to generate suitable BBoxes to serve the re-ID; however, most existing works generate low-quality BBoxes when facing challenges, such as occlusion. Our solution is to construct an SPG with a novel node representation to model the interplay between the pedestrian BBoxes and scene information.

1) Graph Construction: To perform graph feature learning for the scene information, we first need to construct graph $G(V, E)$. We define $V = \{V_p, V_s\}$, where V_p and V_s represent pedestrian and scene nodes, respectively. For V_p, $V_p = \{v_{p_1}^1, \ldots, v_{p_1}^{n_1}, \ldots, v_{p_i}^1, \ldots, v_{p_i}^{n_i}\}$, $v_{p_i}^{n_i}$ represents the node constituted by the ith pedestrian's n_ith BBox, containing the BBox features with latest correction, the IoU with the target pedestrian, the pedestrian ID, and the corresponding temporal information (e.g., video clip number and frame number of the image where the pedestrian appears). For V_s, using a whole scene image as a node can make it difficult to describe and learn the scene information, as shown in Fig. 1. Thus, we divide each scene image into blocks and each block serves as a node. We define $V_s = \{v_{s_1}^1, \ldots, v_{s_1}^k, \ldots, v_{s_j}^1, \ldots, v_{s_j}^k\}$, where $v_{s_j}^k$ represents the node constituted by the kth block of the jth scene, and $v_{s_j}^k \in \mathbb{R}^{W \times H}$. Each scene block node comprises the block location and the confidence of the pedestrian BBoxes located there (with center points falling within the block), to measure their plausibility. Then we define the edge set E which consists of three subsets, E_{ps}, E_{pp}, and E_{ss}, that connect different types of nodes. E_{ps} connects pedestrian and scene nodes, E_{pp} connects pedestrian and pedestrian nodes, and E_{ss} connects scene and scene nodes. For each subset, we define their edges. For E_{ps}, a nondirectional edge, $e_{ps} =< v_{p_i}^{n_i}, v_{s_j}^k >$, where $e_{ps} \in E_{ps}$, indicates that the center point of the ith pedestrian's n_ith BBox is within the kth scene block of the jth scene, and the weight of e_{ps} is calculated as the overlapping area between the BBox and the block. For E_{pp}, a nondirectional edge, $e_{pp} =< v_{p_i}^n, v_{p_i}^m >$, where $e_{pp} \in E_{pp}$, indicates that the two pedestrian nodes have either the same identity ID or highly similar features, and the weight of e_{pp} is their similarity or -1 if there is no similarity. For E_{ss}, a nondirectional and nonweighted edge, $e_{ss} =< v_{s_j}^a, v_{s_j}^b >$, where $e_{ss} \in E_{ss}$, indicates that two blocks are from the same scene and are spatially contiguous. To provide a more intuitive illustration of the node representation and graph construction, we present a schematic sketch of our SPG modeling process in Fig. 4. Clearly, the scene nodes model the various regions of scenes and act as intermediate stations between pedestrian nodes.

2) Adaptive Scene Image Block Size: A scene block should have an appropriate height H and width W to ensure it contains

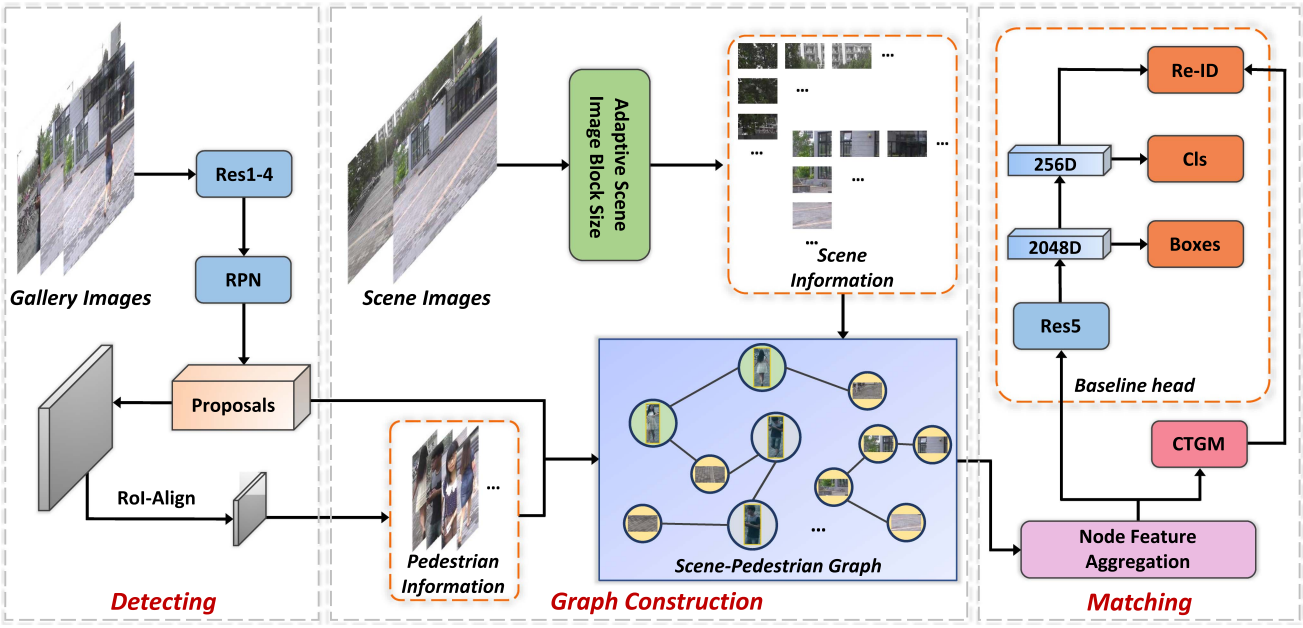

Fig. 3. Overall architecture of the end-to-end framework with our SPG and CTGM. Our network is based on faster R-CNN with a Res2Net50. The framework includes three modules: feature extraction network, scene-pedestrian graph module, and Re-ID network.

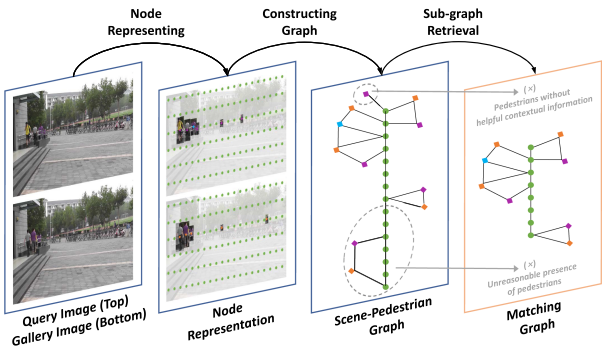

Fig. 4. Schematic diagram of the proposed SPG modeling process. Green circular nodes represent scene nodes, while diamond-shaped nodes represent pedestrian nodes (with blue, purple, and orange corresponding to the query person, other pedestrians in the query image, and pedestrians in the gallery image, respectively).

enough pedestrians to produce discriminative confidence scores between blocks, while avoiding redundant nodes that result in an oversized GCN. In addition, the height and width of the scene image should be divisible by the height and width of the scene blocks to maintain a consistent block size. To achieve this, we propose to find suitable block sizes for each scene and the sizes of the scene blocks are set as follows:

$$\begin{aligned} H &= n_H \cdot h, H_{\text{pmin}} \le h, 1 \le n_H \le \left\lfloor \frac{H_{\text{pmax}}}{h} \right\rfloor \\ W &= n_W \cdot w, W_{\text{pmin}} \le w, 1 \le n_W \le \left\lfloor \frac{W_{\text{pmax}}}{w} \right\rfloor \end{aligned} \tag{1}$$

where H_{pmin} and W_{pmin} represent the height and width of the smallest pedestrian BBox appearing in the ground truth, H_{pmax} and W_{pmax} are the maximum values, h and w denote the smallest factors of the scene image's height and width, respectively, and n_H and n_W are integers starting from 1. Generally, the middle values of n_H and n_W are sufficient.

3) Node Feature Aggregation: After constructing the initial graph structure, it is necessary to determine the most appropriate node feature aggregation rules to convey information. We have conducted an ablation study on our constructed graph, testing three popular GCNs to explore their effects on the performance of our model. The Geom-GCN [25] is applied to our final model, which achieves the best performance. Specifically, the overall node feature aggregation rule is given as follows:

$$h_v^{l+1} = \sigma\left(\boldsymbol{W}_l \cdot \|_{\alpha \in V} \boldsymbol{f}_{(\alpha,r)}^{v,l+1}\right) \tag{2}$$

where $f_{(\alpha,r)}^{v,l+1}$ donates the features of the virtual node generated from node $v \in V$ and indexed by (α, r), (α, r) corresponds to the combination of a neighborhood α and a relationship r (consists of the relevant edges in our SPG), l represents the layer of the GCN, $\boldsymbol{W}_l$ is the weight matrix estimated by backpropagation, and ReLU is used as the nonlinear activation function $\sigma(\cdot)$.

4) Joint Detection and GCN for Person Search: In the detection phase, our SPG can refine the quality of pedestrian BBoxes by using the high-quality BBoxes to guide the low-quality ones in the same scene. Specifically, we adopt BBA [9] as the basic strategy for refining BBoxes by moving the top, bottom, left, and right boundaries of each BBox. As shown in Fig. 5, this strategy may fail to generate accurate corrections when facing complex scenarios, resulting in meaningless or even noisy features in the occluded regions. We assume that high-quality BBoxes rarely or even do not include such meaningless features, i.e., compared to low-quality BBoxes, high-quality ones exhibit greater similarity between local and global features. Therefore, we measure the quality of BBoxes by evaluating the differences

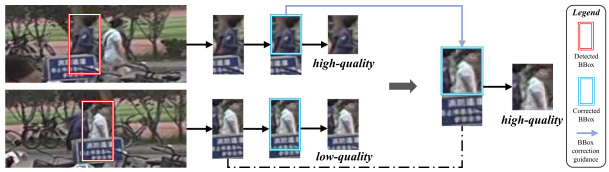

Fig. 5. Illustration of pedestrian BBox correction conducted by SPG.

$D(f_g, f_l)$ between global (stripe) features f_g and local features f_l. For each pedestrian node, we determine whether its $D(f_g, f_l)$ is significantly higher than the average $D(f_g, f_l)$ of pedestrian neighbors connected to the same scene node (based on empirical experience, we set the threshold to 1.75 times the average). If it is, the correction guidance will be activated, and the most frequently repeated BBox correction among the pedestrian neighbors will be copied to the current pedestrian node.

To mitigate the negative effects of unlabeled persons and faulty detection, we assign different confidence distributions to each scene. Specifically, for a scene with a tailored image block size, we first count the pedestrians associated with each block (i.e., the pedestrian neighbors for each scene node, denoted as $N_p(v_{s_j}^k)$, $v_{s_j}^k \in V_s$). Then, for a scene node $v_{s_j}^k$, we determine its total pedestrian neighbors with all scene neighbors $N_s(v_{s_j}^k)$ and calculate their average, denoted as c_j^k. Finally, the values from the previous step are mapped from the real-valued range to the decimal space with a range of $[\gamma, 1.00]$. The final results are used as the confidence levels $\hat{c}$ of the corresponding scene blocks, formulated as

$$c_j^k = \frac{1}{\left|N_s\left(v_{s_j}^k\right)\right|} \sum_{v_{s_i}^k \in N_s\left(v_{s_j}^k\right)} TN_p\left(v_{s_j}^k, v_{s_i}^k\right) \tag{3}$$

$$\hat{c}_j^k = \gamma + (1-\gamma) \times \frac{c_j^k - \min_i c_i^k}{\max_i c_i^k - \min_i c_i^k} \tag{4}$$

where γ is a scale factor and is set to 0.96 according to the experimental results. $TN_p(v_{s_j}^k, v_{s_i}^k)$ denotes the total number of unique pedestrian neighbors of $v_{s_j}^k$ and $v_{s_i}^k$, which can be calculated as $|N_p(v_{s_j}^k) \cup N_p(v_{s_i}^k)|$. In this way, by smoothing the confidence distribution, SPG can reduce the priority of unreasonable BBoxes without affecting the reasonable ones. To improve matching efficiency, as shown on the right side of Fig. 4, pedestrian nodes with unreasonable BBoxes will not be included in the matching graph.

To leverage GCN, we apply two baseline heads to features obtained in GCN layers, which have superior features after node feature aggregation by encoding relations. The first head's regression loss L_{reg_1} and classification loss L_{cls_1} are defined as follows:

$$L_{\text{reg}_1} = \frac{1}{N_p} \sum_{i=1}^{N_p} L_{\text{loc}}\left(r_i, \Delta_i\right) \tag{5}$$

$$L_{\text{cls}_1} = -\frac{1}{N} \sum_{i=1}^{N} c_i \log\left(p_i\right) \tag{6}$$

where N_p is the number of positive samples, r_i is the calculated regressor of the ith positive sample, $\triangle_i$ is the corresponding ground truth regressor, p_i is the predicted classification probability of the ith sample, c_i is the ground truth label, and L_{loc} is the smooth-L_1-loss.

For the second head, it shares the same regression loss as the first head. Its classification and re-ID losses are the NAE $L_{\text{nae}}(.)$ [15]. The overall loss of our network is the sum of the two heads' losses over all layers of GCN

$$L_{\text{total}} = \sum_l \eta_1 L_{\text{reg}_1}^l + \eta_2 \left(L_{\text{cls}_1}^l + L_{\text{reg}_2}^l + L_{\text{cls}_2}^l + L_{re-ID}^l\right) \tag{7}$$

where l is the index of GCN layers and η_1, η_2 denote weights of each loss. We adopt an automatic loss weighting scheme [26] to balance η_1 and η_2.

D. Contextual and Temporal Graph Matching

To reduce the mismatching caused by similar-looking pedestrians and further refine re-ID, we design a CTGM algorithm to exploit the contextual and temporal information presented in the constructed graph for pedestrian matching.

1) Unveiling the Contextual and Temporal Information: For the contextual information, affected by problems, such as cross-camera view differences and occlusion, a single-point matching strategy (i.e., only focusing on the similarity between the query pedestrian and the candidate pedestrian) may be unreliable. To mitigate this, we introduce contextual information by weighting the similarity of surrounding pedestrians to obtain more reliable matching results. For instance, in the left side of Fig. 6, the correct match for pedestrian (a) should be pedestrian (b), but due to (b) being obscured by the other pedestrian, the similarity between (b) and (a) is lower than the similarity between (c) and (a). After applying the CTGM algorithm to calculate the similarity of the surrounding pedestrians between (a) and (b) and (a) and (c), respectively, the correct match (b) is obtained.

In terms of the temporal information, we observe that pedestrians are mostly walking posture at a slow speed and within a scene, the same pedestrian tends to appear in a certain time period. Taking the right side of Fig. 6 as an example, the correct match (f) has a lower similarity to (e) than (g), due to the interference of the white-clad pedestrian located next to (f). If we can reweight the matching score with temporal relationship (i.e., pay more attention to the proximity of pedestrians while considering their spatial distance), the weight of the incorrect match (g) can be reduced, leading to the correct match (f). To model the temporal information discussed above, the data of cameras, segments, and frames provided by datasets is encoded into the pedestrian nodes (e.g., "62056793" denotes the temporal data of camera/scene 06, segment 02, and frame 056793 for PRW), and these are utilized in the pedestrian matching process. Specifically, when we find that the query pedestrian and the candidate pedestrian are located in the same scene, we retrieve their temporal data and check if they are temporally adjacent (i.e., appear in the same scene/camera and segment within 1000 frames). If they are, the matching score can be appropriately

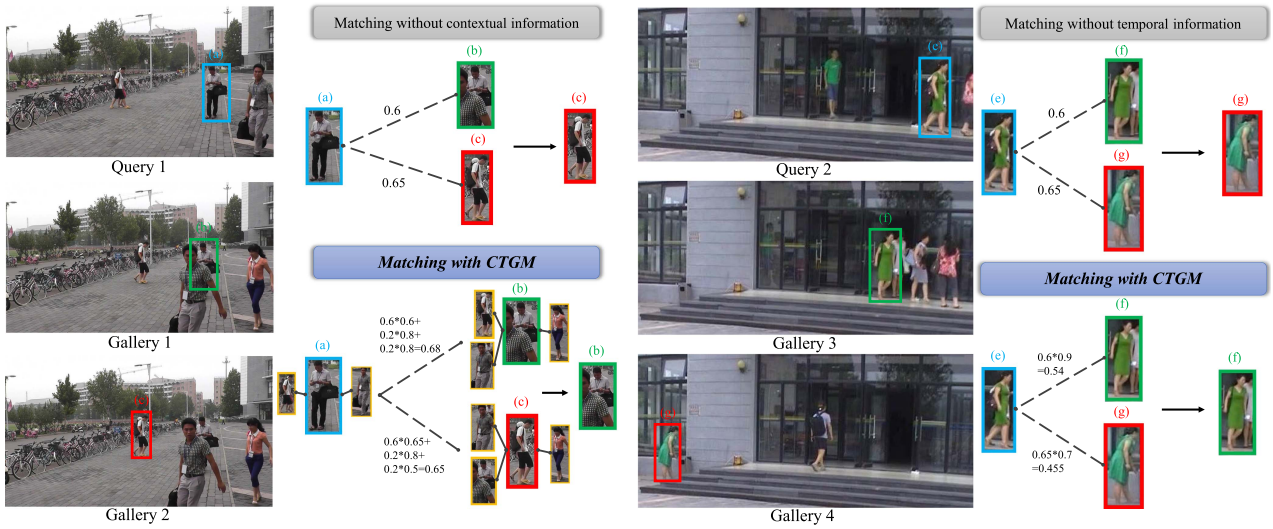

Fig. 6. Two examples of matching query images with gallery images, with the values attached to each dashed line representing the similarity between the corresponding pedestrians. Notably, the incorporation of contextual and temporal information effectively prevents false matches in specific cases.

reduced when two temporally adjacent pedestrians are too far apart.

2) Formulating the Pedestrian Matching Process With CTGM: Sharing with previous works [16] and taking Fig. 6 for example, we define the following symbols.

1) Q/G: The query/gallery image.
2) q: The query pedestrian in Q [the blue boxes, i.e., person (a) and (e)]
3) P: All pedestrians in an image ($\{(a),(e)\} \in P_Q$)
   $\text{sim}(p_1, p_2)$: The cosine similarity between pedestrian p_1 and p_2 calculated by extracted features.
   $\text{SIM}(q, G)$: The similarity between q and G, defined as the maximum value among these similarities between q and P_G

$$\text{SIM}(q, G) = \max_{p \in P_G} \text{sim}(q, p). \tag{8}$$

4) $M(\mathbb{V}, \mathbb{E})$: The matching graph. It is a subgraph of SPG and consists of the relevant pedestrian and scene nodes involved in matching, used to model the contextual and temporal relationship between q and P_G.
5) $W_t(v_i, v_j)$: The weight of the temporal information between the pedestrian nodes $v_i, v_j \in \mathbb{V}_p$, calculated from the distance between the pedestrians as follows:

$$\begin{aligned} S(v_i) &= \max_{v_k \in \mathbb{T}(v_i)} dist\left(v_i, v_k\right) \\ D\left(v_i, v_j\right) &= \begin{cases} \frac{\max(S(v_i), S(v_j))}{dist(v_i, v_j)}, & \mathbb{T}\left(v_i\right) \cup \mathbb{T}\left(v_j\right) ! = \varnothing \\ 1, & \text{else.} \end{cases} \\ W_t\left(v_i, v_j\right) &= \begin{cases} n_t D\left(v_i, v_j\right), & n_t D\left(v_i, v_j\right) < 1 \\ 1, & \text{else.} \end{cases} \end{aligned} \tag{9}$$

where $\text{dist}(v_i, v_j)$ denotes the spatial distance between the BBoxes of nodes v_i, v_j, $\mathbb{T}(v_i)$ denotes the set of nodes that are temporally adjacent to v_i and have the same pedestrian ID, and n_t is a temporal multiplication parameter studied in ablation experiments.

6) $W_c(v_i, v_j)$: The weight of the contextual information between the pedestrian nodes $v_i, v_j \in \mathbb{V}_p$. It is defined as the sum of the related pedestrians' similarities

$$W_c\left(v_i, v_j\right) = \sum_{\substack{(v_i, v_l) \in \mathbb{E}_{pp}, \\ (v_j, v_k) \in \mathbb{E}_{pp}, \\ (v_l, v_k) \in \mathbb{E}_{pp}}} \text{sim}\left(p_l, p_k\right). \tag{10}$$

7) $C(M)$: The final confidence of the matching graph M, calculated by weighting the temporal and contextual information on the basis of similarity as follows:

$$C(M) = \max_{v_i \in \mathbb{V}} \left[\lambda_1 W_t\left(q, v_i\right) \cdot \text{sim}\left(q, v_i\right) + \lambda_2 W_c\left(q, v_i\right)\right]. \tag{11}$$

We describe the proposed CTGM algorithm in Algorithm 1. All the gallery images are sorted in descending order according to $\text{SIM}(q, G)$ and we keep the top-30 for processing. Moreover, to minimize the noise caused by excessive background information, we only consider pedestrians with top-3 detection confidence.

IV. Experiments

In this section, we first introduce our experimental settings, including datasets, evaluation protocols, and implementation details. Next, we compare our method with the state-of-the-art ones. Then we evaluate the performance of each component through ablation experiments. Finally, we present more performance and qualitative results to further demonstrate the efficacy of the proposed method.

A. Experimental Settings

1) Datasets: In our experiments, we use two large-scale person search benchmarks PRW and CUHK-SYSU. PRW [41]

Algorithm 1: CTGM.

Input:
Query image, Q
Query pedestrian, $q \in P_Q$
Gallery images, $S = \{G_1, G_2, \ldots\}$
Output:
The most similar pedestrian to q in each gallery image,
Similarities between q and these most similar pedestrians

```
Rank S in descending order by SIM(q, G)
Remain top-30 gallery images, S = {G_1, G_2, ..., G_{k_1}}
Rank P_Q in descending order by detection confidence
Remain top-3 pedestrians, P_Q = {q_1, q_2, ..., q_{k_2}}
Set pedestrians, sims to empty list
for each G ∈ S do
    Based on P_Q and P_G, get its matching graph M(V, E)
    for each vertices v_i of M do
        if v_i = q then
            Insert all v_j that satisfy (v_i, v_j) ∈ E into pedestrians
            Insert C(M) into sims
            break
        end if
    end for
end for
return pedestrians, sims
```

is a widely used dataset with 11 816 scene images captured from six cameras on a university campus. 34 304 pedestrian bounding boxes of 932 pedestrians are manually annotated. The training set contains 5134 images with 482 different pedestrians, while the testing set contains 6112 images with 2057 query pedestrians.

CUHK-SYSU [11] is a large-scale dataset consisting of 18 184 scene images, mainly from street cameras and movie screenshots with a total of 96 143 labeled pedestrians and 8432 pedestrian IDs. The training set contains 11 206 scene images with 5532 pedestrians and 55 272 annotated BBoxes. The test set consists of 6978 scene images with 2900 pedestrians, and the number of labeled pedestrians reaches 40 871.

2) Evaluation Protocols: We utilize the same evaluation protocols as in previous works [16], where the mean average precision (mAP) and the cumulative matching characteristic (CMC) are adopted as evaluation metrics.

3) Implementation Details: Our model is implemented with PyTorch, running on one NVIDIA RTX 3090 GPU. During the training phase, we resize each image to 900 $\times$ 1500 pixels and the batch size is 5. Our model is optimized by SGD and the final hyperparameter setting is an initial learning rate of 0.03, the patience of 20 epochs (18 for PRW), the SGD momentum and weight decay of 0.9 and 5×10^{-4} individually. The circular queue size of OIM is set to 5000/500 for CUHK-SYSU/PRW and we use nonmaximum suppression (NMS) with a 0.4/0.5 threshold in the test phase to remove redundant BBoxes.

B. Comparison With the State-of-the-Art Methods

In this section, we compare our method with the state-of-the-art models on CUHK-SYSU and PRW.

1) Performance Comparison on PRW: The left column of Table I demonstrates the superiority of our method in terms of rank-1 accuracy over other competitors on PRW. Compared to our baseline SeqNet, our method yields a 1.79 % and 6.44% improvement in mAP and rank-1, respectively. When compared with the state-of-the-art two-stage model SEIE [31], our SPG+CTGM outperforms it by 3.24% w.r.t rank-1. In addition, our method surpasses the mAP and rank-1 accuracy of the state-of-the-art end-to-end models COAT [40] and COAT+CBGM [40] by 2.44% and 0.74% w.r.t rank-1, respectively, though COAT adopts transformers with heavier computational burden. The PRW dataset offers an adequate number of samples for each scene, enabling the SPG to gather more useful information and better its construction. These results confirms the great potential of our approach in practical applications, as real-world data can provide a more diverse set of samples to further enhance the effectiveness of our SPG.

2) Performance Comparison on CUHK-SYSU: The right column of Table I reports that our method outperforms our baseline SeqNet by 1.26% and 1.35% w.r.t mAP and rank-1 on CUHK-SYSU. In addition, our approach surpasses both the best two-stage and end-to-end methods, indicating the effectiveness of SPG+CTGM in various scenarios. Compared to PRW, CUHK-SYSU provides fewer images for each scene and lacks temporal information. Thus, the SPG constructed on CUHK-SYSU is more scattered and the learned interplay between scenes and pedestrians is not as adequate as on PRW.

3) Universality of SPG and CTGM: We further apply our approach to other end-to-end frameworks including OIM [11], HOIM [14], and NAE [15]. As shown in Table I, our method can effectively enhance the performance of OIM, HOIM, NAE, and SeqNet, in particular, the improvement of OIM on PRW is significant (mAP $\uparrow$ 10.63% and rank-1 $\uparrow$ 9.35%).

C. Effectiveness of GCNs

In order to maximize the effectiveness of our method, we choose the most suitable GCN and determine the optimal number of layers.

1) Selection of GCN: We have adopted three popular GCNs on PRW and CUHK-SYSU, including GraphConv [42], Geom-GCN [25], and COMPGCN [43]. We apply the optimal scene image block size and use three layers. As shown in Fig. 7(a), Geom-GCN provides the best performance among the three, leading to further improvement compared to the baseline SeqNet (mAP $\uparrow$ 0.93% and rank-1 $\uparrow$ 1.02% on CUHK-SYSU, mAP $\uparrow$ 0.65% and rank-1 $\uparrow$ 3.26% on PRW). This outcome is expected since Geom-GCN's virtual node design enables it to better capture the relationships among neighbors. Our experiments show that Geom-GCN is more suitable for datasets with low homogeneity, which aligns with the feature that the person search datasets have many different pedestrians and few samples of the same pedestrian.

TABLE I
PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART METHODS ON CUHK-SYSU AND PRW

	Method	Additional Information Required for Re-ID	PRW		CUHK-SYSU	
			mAP	Rank-1	mAP	Rank-1
two-stage	MGTS(ECCV18) [27]	Mask	32.60	72.10	83.00	83.70
	RDLR(ICCV'19) [8]	None	42.90	70.20	93.00	94.20
	IGPN(CVPR'20) [28]	None	47.20	87.00	90.30	91.40
	TCTS(CVPR'20) [29]	None	46.80	87.50	93.90	95.10
	Faster R-CNN+PCB+OR(TIP'20) [10]	None	43.00	65.90	92.90	93.70
	BTCL(TMM'22) [30]	None	47.40	**88.30**	94.20	95.60
	SEIE(TCSVT'22) [31]	None	**54.00**	86.70	**95.00**	**95.80**
end-to-end	OIM(CVPR'17) [11]	None	21.30	49.90	75.50	78.70
	RCAA(ECCV'18) [32]	None	-	-	79.30	81.30
	CTXGraph(ICCV'19) [23]	None	33.40	73.60	84.10	86.50
	CA-MN(TC'19) [33]	Knowledge Distillation	34.50	59.90	91.10	91.90
	HOIM(CVPR'20) [14]	None	39.80	80.40	89.70	90.80
	BPNet(AAAI'20) [34]	Keypoint&Mask	48.50	87.90	88.40	90.50
	BINet(CVPR'20) [35]	Knowledge Distillation	45.30	81.70	90.00	90.70
	APNet(CVPR'20) [9]	Keypoint	41.90	81.40	88.90	89.30
	NAE(CVPR'20) [15]	Detection Confidence	43.30	80.90	91.50	92.40
	AlignPS+(CVPR'21) [17]	None	46.10	82.10	94.00	94.50
	SeqNet(AAAI'21) [16]	Detection Confidence	46.70	83.40	93.80	94.60
	SeqNet+CBGM(AAAI'21) [16]	Detection Confidence	47.60	87.60	94.80	95.70
	DKD(AAAI'21) [36]	Knowledge Distillation	50.50	87.10	93.10	94.20
	CANR+(TCSVT'22) [37]	None	44.80	83.90	93.90	94.50
	BUFF(TCSVT'22) [38]	None	44.90	86.30	91.60	92.20
	DMRNet++(TPAMI'22) [39]	None	52.10	87.00	94.50	95.70
	COAT(CVPR'22) [40]	Transformer	53.30	87.40	94.20	94.70
	COAT+CBGM(CVPR'22) [40]	Transformer	**54.00**	89.10	94.80	95.20
	OIM+SPG+CTGM(ours)	None	44.63(10.63 ↑)	85.25(9.35 ↑)	91.24(4.14 ↑)	90.92(2.42 ↑)
	HOIM+SPG+CTGM(ours)	None	43.83(4.03 ↑)	85.36(4.96 ↑)	91.65(1.95 ↑)	92.44(1.64 ↑)
	NAE+SPG+CTGM(ours)	Detection Confidence	45.77(2.47 ↑)	87.71(6.81 ↑)	92.58(1.08 ↑)	93.16(0.76 ↑)
	SeqNet+SPG+CTGM(ours)	Detection Confidence	48.49(1.79 ↑)	**89.84**(6.44 ↑)	**95.06**(1.26 ↑)	**95.95**(1.35 ↑)

Our models are shown in italics.
The bold entities indicate the best performance achieved by two-stage and end-to-end methods, separately.

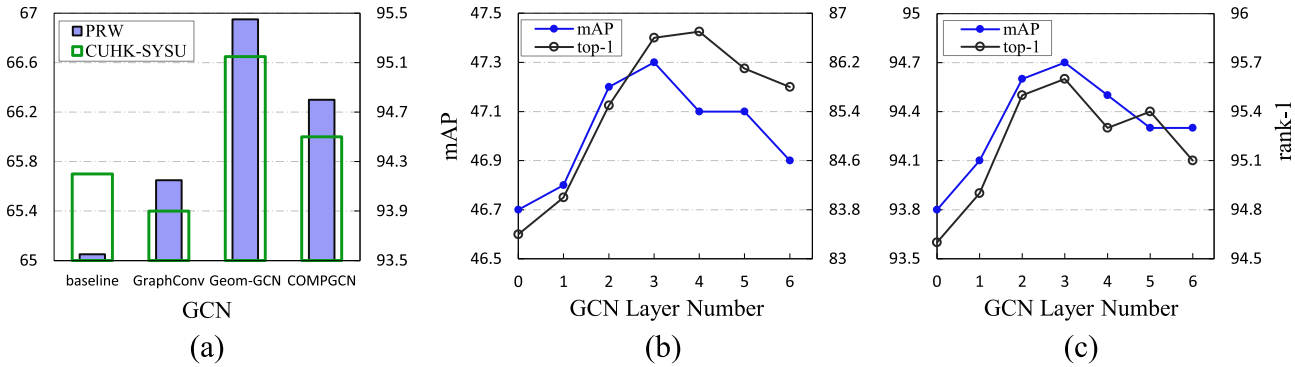

Fig. 7. (a) Performance comparison of SPG with different GCNs on PRW (the primary *y*-axis) and CUHK-SYSU (the secondary *y*-axis) evaluated by $\frac{\text{mAP}+\text{rank}-1}{2}$. (b) and (c) Effects of the GCN layer number evaluated on (b) PRW and (c) CUHK-SYSU. We choose Geom-GCN and set the optimal GCN layer number as 3.

2) Optimal GCN Layer Number: Intuitively, increasing the number of GCN layers is expected to enhance contextual information by allowing nodes to aggregate features from a wider range of neighbors. However, an excessive number of layers may lead to cluttered node features and memory overhead. To find the optimal number of layers, we compare the performance of SPG with {1,2,3,4,5,6}-layers GCNs in Fig. 7(b) and (c). Our results demonstrate that the 3- and 4-layer GCNs outperform the others. To balance the effectiveness and computation cost, we set the optimal number of GCN layers as 3.

3) Complexity Analysis: It is significant to provide a complexity analysis of our SPG. In theory, based on Geom-GCN [25] with the optimal parameters discussed above, the computational complexity of SPG is $O(f \times 2|n| \times m)$ to update the representations of one node, where f donates the input representation size (i.e., 256-d pedestrian features), n and m donate the number of virtual nodes and hidden units for each virtual node (i.e., (α, r)), respectively. Our approach employs 16 hidden units with 8 virtual nodes. In addition, the real running time of our approach can be found in Table IV.

D. Ablation Study

In this section, we perform detailed ablation experiments on two datasets to evaluate the effectiveness of the proposed SPG with CTGM and better analyze each component of our method.

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT DETECTORS AND REIDENTIFIERS ON PRW AND CUHK-SYSU

	Detector	Recall	AP	Re-identifier	mAP	rank-1
PRW	SeqNet	96.7	94.2	SeqNet	46.70	83.40
				SPG+CTGM	48.03	88.54
	SPG	98.1	96.0	SeqNet	47.32	86.64
				SPG+CTGM	48.49	89.84
	GT	100	100	SeqNet	48.12	88.27
				SPG+CTGM	49.64	89.95
CUHK-SYSU	SeqNet	92.1	89.2	SeqNet	93.80	94.50
				SPG+CTGM	94.92	95.77
	SPG	92.5	90.3	SeqNet	94.14	95.09
				SPG+CTGM	95.06	95.95
	GT	100	100	SeqNet	94.60	95.30
				SPG+CTGM	95.67	96.46

1) Comparison With Different Detectors and Re-Identifiers: To investigate the performance improvement brought by the proposed SPG and CTGM separately, we adopt different detectors in the pedestrian detection stage and various re-identifiers in the re-ID stage, and summarize the results in Table II. The left side of Table II shows that SPG outperforms the baseline SeqNet in both AP and recall metrics on both datasets, indicating that our SPG achieves better detection than SeqNet. This results in higher quality pedestrian BBoxes for the re-ID stage and better performance, as shown on the right side of Table II. The reason for this is that our SPG can remove the wrong pedestrian BBoxes appearing in the unreasonable areas and conduct accurate pedestrian BBox corrections by using the high-quality BBoxes to guide the low-quality ones in the same scene.

Furthermore, as reported in the right side of Table II, when using the same SeqNet detector for fair comparisons, our SPG with CTGM can boost rank-1 by 5.14% and 1.27% on PRW and CUHK-SYSU, respectively. Similar improvement (3.20%↑ on PRW, 0.86%↑ on CUHK-SYSU) can be observed when adopting the SPG detector. These comparisons demonstrate the effectiveness of our CTGM in various scenarios on both datasets. To push the limit of our method, we only focus on the re-ID stage and further adopt ground truth BBoxes as detection results. Finally, based on our proposed SPG and CTGM, the performance achieves the mAP of 95.67% and 49.64% on the CUHK-SYSU and PRW datasets, respectively.

2) Different n_t of CTGM: We evaluate the performance with different temporal multiplication parameter n_t of CTGM on PRW. Table III shows that CTGM achieves the best performance when n_t is set to 0.9. In addition, we observe that excessively low values of n_t result in more fluctuations in performance when compared to higher values. This can be attributed to the increased impact of temporal information on model predictions when n_t is lower, i.e., leading to a higher number of affected pedestrian matches. However, it is important to note that the temporal information should aid the model in making appropriate corrections without overcompensating.

TABLE III
ABLATION STUDY OF TEMPORAL MULTIPLICATION PARAMETER n_t CONDUCTED ON PRW

n_t	mAP+rank-1	matches affected
0.70	133.24	1.2%
0.75	134.73	2.5%
0.80	135.87	3.1%
0.85	136.29	4.4%
0.90	**136.47**	**4.7%**
0.95	136.38	4.9%
1.00	135.71	5.3%

The bold entities indicate the best results measured by mAP+rank-1.

TABLE IV
ABLATION STUDY ABOUT DIFFERENT COMPONENTS OF PERFORMANCE AND SPEED EVALUATED ON PRW

SPG	CGM	TGM	mAP	rank-1	time cost (ms)
✗	✗	✗	46.70	83.40	361
✓	✗	✗	47.32	86.64	448
✓	✓	✗	48.22	88.73	450
✓	✗	✓	47.95	88.56	451
✓	✓	✓	**48.49**	**89.84**	454

The bold entities indicate the best results measured by mAP and rank-1.

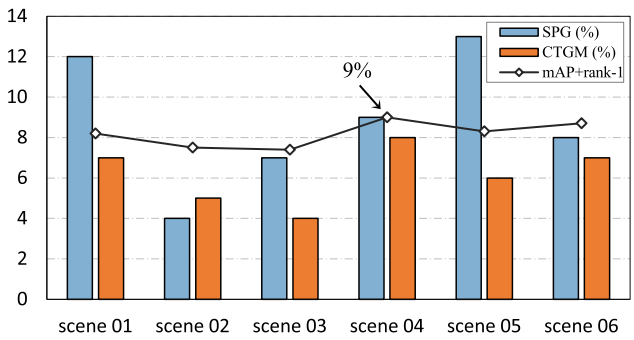

Fig. 8. Details of the effects from SPG and CTGM. SPG (%) denotes the percentage of pedestrian BBoxes corrected in SPG; CTGM (%) denotes the percentage of matches changed by CTGM; mAP+rank-1 represents the performance boosting for each scene evaluated by the improvement on metric (mAP + rank − 1); *x*-axis denotes the six scenes in PRW.

3) Effectiveness of the SPG and CTGM Components: The last three rows of Table IV show that the contextual and temporal information is effectively utilized by CTGM with 3-layers GCN. The use of contextual information (CGM) results in a 0.90% increase in mAP and a 2.09% increase in rank-1, while the use of temporal information shows similar improvement (mAP ↑ 0.63% and rank-1 ↑ 1.92%). Notably, the additional time cost of computation brought by CTGM is light (about 6 ms) according to the last column. To further explore the contribution of SPG and CTGM in more detail, we also calculate the percentage of pedestrian BBoxes corrected by SPG and the number of matching pairs affected by CTGM in the six scenes of PRW. As shown in Fig. 8, the case of scene 05, which is most affected by CTGM (the highest orange pillar), achieves the most significant improvement of 9% measured by (mAP

TABLE V
DETAILED RESULTS OF THE ADAPTIVE IMAGE BLOCK SIZES FOR VARIOUS SCENES ON PRW

Scene ID	Frame Scale	$W_{\text{pmin}} \times H_{\text{pmin}}$	$W_{\text{pmax}} \times H_{\text{pmax}}$	$w \times h$	n_W	n_H	Optimal Block Size
01	1920×1080	25×58	574×777	30×60	{1, 2, 4, 8, 16}	{1, 2, 3, 6, 9}	120×180
02	1920×1080	30×67	247×460	30×72	{1, 2, 4, 8}	{1, 3, 5}	60×216
03	1920×1080	21×58	321×623	24×60	{1, 2, 4, 8, 10}	{1, 2, 3, 6, 9}	96×180
04	1920×1080	37×79	531×717	40×90	{1, 2, 4 ,6, 8, 12}	{1, 2, 3, 4, 6}	80×270
05	1920×1080	31×81	393×704	32×90	{1, 2, 3, 4, 6, 10}	{1, 2, 3, 4, 6}	128×180
06	720×576	23×71	233×461	24×72	{1, 2, 3, 5,6}	{1, 2, 4}	72×144

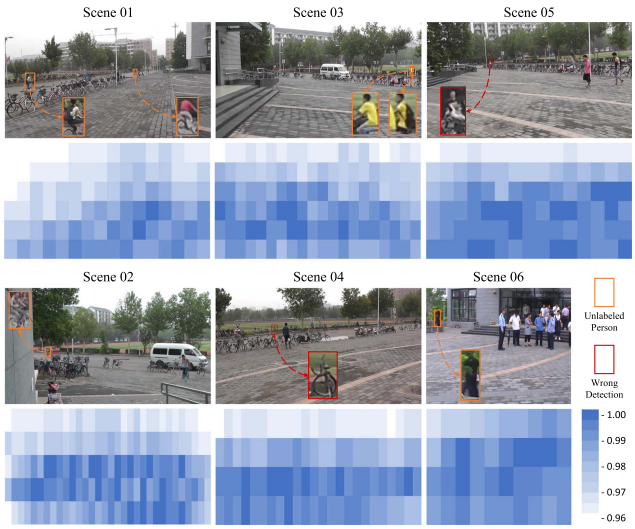

Fig. 9. Visualization of the pedestrian BBox confidence distribution on 6 scenes of PRW. We use blue/white for each image block to represent the reasonable/unreasonable presence of pedestrians at that location and show several cases (unlabeled persons and wrong detections) of correctly demoted priorities.

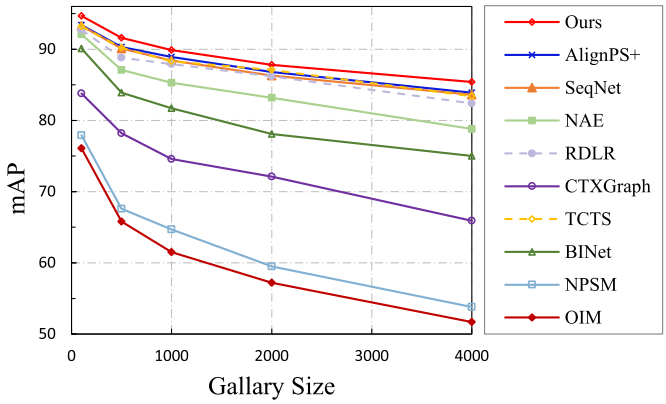

Fig. 10. Comparative results on CUHK-SYSU with different gallery sizes. The solid and dashed lines represent end-to-end and two-stage methods, respectively.

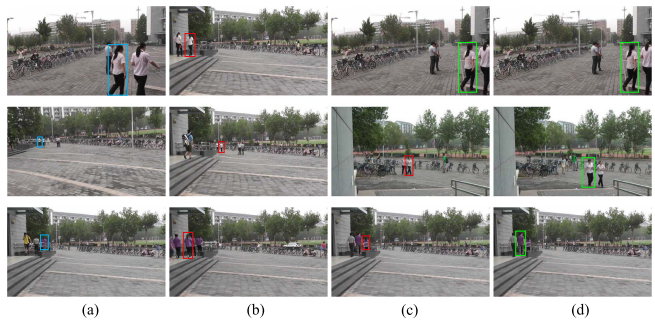

Fig. 11. Visualization of rank-1 results on PRW dataset and each row of images is a group. The blue bounding boxes denote the queries, while the green and red bounding boxes represent the correct and incorrect rank-1 matches, respectively.

Fig. 12. Visualization of rank-1 results on CUHK-SYSU dataset and each row of images is a group. The blue bounding boxes denote the queries, while the green and red bounding boxes represent the correct and incorrect rank-1 matches, respectively.

+ rank-1). This indicates that the the performance bottleneck of person search networks does not lie in the detection stage and that the re-ID stage breakthrough, such as the multipoint matching strategy offered by CTGM, can yield more substantial performance gains when the performance approaches saturation.

4) SPG Provides Targeted Solutions for Each Scene: Our SPG can tailor adaptive image block sizes for each scene to maximize the model's performance. In Table V, we report the detailed results of the adaptive image block sizes for various scenes on PRW and adopt the optimal block size for each scene in the performance experiments. In addition, we visualize the distribution of pedestrian BBox confidence on all scenes in Fig. 9 and provide several examples of the unlabeled persons and wrong detection which have been correctly assigned with lower confidence. It can be noticed that the blue areas are concentrated in reasonable locations where pedestrians often appear, such as road surfaces and building entrances, whereas the unreasonable locations, such as trees, sky, and the edges of the scenes are assigned relatively lower confidence levels. These small distinctions reorder some of the wrong detection BBoxes and bring improvement in detection accuracy as well as the pedestrian matching performance. Moreover, we strive to maintain continuity in the confidence level variations in the

reasonable region, to avoid the negative impact on the correct BBoxes caused by mutations.

E. More Performance and Qualitative Experiments

1) Performance Comparison Under Different Gallery Sizes: To evaluate the robustness of our method under different gallery sizes, we also visualize the results w.r.t mAP with the increase of gallery size on CUHK-SYSU, compared with various end-to-end and two-stage models. Fig. 10 shows the detailed results, indicating that our method outperforms all the other models by notable margins, in terms of all the gallery sizes.

2) Qualitative Experiments on PRW and CUHK-SYSU: We visualize the retrieving results and some qualitative examples are shown in Figs. 11 and 12. Rank-1 person search matches on both PRW and CUHK-SYSU are reported. Compared with the baseline method SeqNet and the classical method OIM, our method is more robust in handling scale/viewpoint variations and is more effective for crowded scenarios, leading to the best performance with the correct matching ranked at the top.

V. Conclusion

In this article, we uncover valuable information in scenes and construct an SPG through a novel node representation to generate higher quality pedestrian BBoxes in end-to-end frameworks. Moreover, we design a CTGM algorithm to enhance the pedestrian matching robustness by utilizing the contextual and temporal information present in the constructed graph. We conducted extensive analytical experiments with ablation study and the experimental results demonstrated that our SPG and CTGM can significantly improve the performance of previous end-to-end models at an acceptable time cost. Our proposed method achieves better performance than the latest deep-learning models with regard to two metrics (mAP and rank-1) on both benchmarks PRW and CUHY-SYSU.

Limitations and Future Work: While the proposed approach performs well in general, our method requires temporal information (e.g., video clip number and frame number) provided by the dataset to activate the TGM algorithm, enabling better matching of walking pedestrians. Consequently, the performance of our method is somewhat restricted when temporal information is unavailable. In the future, we will construct large-scale synthetic datasets for person search and continue to explore the person search task with the help of graph structures in a wider range of application scenarios (e.g., text-based person search).

Ethical Considerations: Person search methods, like most technologies, have the potential to yield both societal benefits and negative consequences. For instance, by identifying target individuals, person search can assist in apprehending suspects and counter-terrorism operations. However, the irresponsible implementation of this technology may invade personal privacy and its usage should be limited to public areas (e.g., malls, airports, and parks). In addition, we strongly advocate for the development of ethical person search datasets and emphasize the need for further research that prioritizes the protection of personal privacy.

References

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

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[3] Y. Yang, L. Wen, S. Lyu, and S. Li, "Unsupervised learning of multi-level descriptors for person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, 2017, pp. 4306–4312.

[4] X. Zang, G. Li, and W. Gao, "Multi-direction and multi-scale pyramid in transformer for video-based pedestrian retrieval," *IEEE Trans. Ind. Inform.*, vol. 18, no. 12, pp. 8776–8785, Dec. 2022.

[5] M. Ye, Y. Cheng, X. Lan, and H. Zhu, "Improving night-time pedestrian retrieval with distribution alignment and contextual distance," *IEEE Trans. Ind. Inform.*, vol. 16, no. 1, pp. 615–624, Jan. 2020.

[6] L. Wu et al., "Pseudo-pair based self-similarity learning for unsupervised person re-identification," *IEEE Trans. Image Process.*, vol. 31, pp. 4803–4816, 2022.

[7] D. Liu et al., "Generative metric learning for adversarially robust open-world person re-identification," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 19, no. 1, pp. 1–19, 2022.

[8] C. Han et al., "Re-ID driven localization refinement for person search," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9814–9823.

[9] Y. Zhong, X. Wang, and S. Zhang, "Robust partial matching for person search in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6827–6835.

[10] H. Yao and C. Xu, "Joint person objectness and repulsion for person search," *IEEE Trans. Image Process.*, vol. 30, pp. 685–696, 2020.

[11] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3415–3424.

[12] C. Gao, R. Yao, J. Zhao, Y. Zhou, F. Hu, and L. Li, "Structure-aware person search with self-attention and online instance aggregation matching," *Neurocomputing*, vol. 369, pp. 29–38, 2019.

[13] B. Munjal, S. Amin, F. Tombari, and F. Galasso, "Query-guided end-to-end person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 811–820.

[14] D. Chen, S. Zhang, W. Ouyang, J. Yang, and B. Schiele, "Hierarchical online instance matching for person search," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 10518–10525.

[15] D. Chen, S. Zhang, J. Yang, and B. Schiele, "Norm-aware embedding for efficient person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12615–12624.

[16] Z. Li and D. Miao, "Sequential end-to-end network for efficient person search," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 2011–2019.

[17] Y. Yan et al., "Anchor-free person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7690–7699.

[18] J. B. Estrach, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and deep locally connected networks on graphs," in *Proc. 2nd Int. Conf. Learn. Representations*, 2014, pp. 1–14.

[19] F. R. Chung and F. C. Graham, *Spectral Graph Theory*. Ann Arbor, MI, USA: American Math. Soc., 1997, no. 92.

[20] Y. Wang, K. Kitani, and X. Weng, "Joint object detection and multiobject tracking with graph neural networks," in *Proc. Int. Conf. Robot. Automat.*, 2021, pp. 13 708–13 715.

[21] K. Guo, Y. Hu, Y. Sun, S. Qian, J. Gao, and B. Yin, "Hierarchical graph convolution networks for traffic forecasting," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 1, 2021, pp. 151–159.

[22] D. Chen, M. Wang, H. Chen, L. Wu, J. Qin, and W. Peng, "Cross-modal retrieval with heterogeneous graph embedding," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 3291–3300.

[23] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu, and X. Yang, "Learning context graph for person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2158–2167.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[25] H. Pei, B. Wei, K. C.-C. Chang, Y. Lei, and B. Yang, "Geom-GCN: Geometric graph convolutional networks," in *Proc. 8th Int. Conf. Learn. Representations*, 2020.

[26] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7482–7491.

[27] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai, "Person search via a mask-guided two-stream CNN model," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 734–750.
[28] W. Dong, Z. Zhang, C. Song, and T. Tan, "Instance guided proposal network for person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2585–2594.
[29] C. Wang, B. Ma, H. Chang, S. Shan, and X. Chen, "TCTS: A task-consistent two-stage framework for person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11952–11961.
[30] C. Wang, B. Ma, H. Chang, S. Shan, and X. Chen, "Person search by a bi-directional task-consistent learning model," *IEEE Trans. Multimedia*, vol. 25, pp. 1190–1203, 2022.
[31] X. Ke, H. Liu, W. Guo, B. Chen, Y. Cai, and W. Chen, "Joint sample enhancement and instance-sensitive feature learning for efficient person search," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7924–7937, Nov. 2022.
[32] X. Chang, P.-Y. Huang, Y.-D. Shen, X. Liang, Y. Yang, and A. G. Hauptmann, "RCAA: Relational context-aware agents for person search," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 84–100.
[33] Y. Zhang, X. Li, and Z. Zhang, "Efficient person search via expert-guided knowledge distillation," *IEEE Trans. Cybern.*, vol. 51, no. 10, pp. 5093–5104, Oct. 2021.
[34] K. Tian, H. Huang, Y. Ye, S. Li, J. Lin, and G. Huang, "End-to-end thorough body perception for person search," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 07, 2020, pp. 12079–12086.
[35] W. Dong, Z. Zhang, C. Song, and T. Tan, "Bi-directional interaction network for person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2839–2848.
[36] X. Zhang, X. Wang, J.-W. Bian, C. Shen, and M. You, "Diverse knowledge distillation for end-to-end person search," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 4, 2021, pp. 3412–3420.
[37] C. Zhao et al., "Context-aware feature learning for noise robust person search," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 7047–7060, Oct. 2022.
[38] W. Yang, H. Huang, X. Chen, and K. Huang, "Bottom-up foreground-aware feature fusion for practical person search," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 262–274, Jan. 2022.
[39] C. Han et al., "DMRNet++: Learning discriminative features with decoupled networks and enriched pairs for one-step person search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7319–7337, Jun. 2023.
[40] R. Yu et al., "Cascade transformers for end-to-end person search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7267–7276.
[41] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1367–1376.
[42] C. Morris et al., "Weisfeiler and leman go neural: Higher-order graph neural networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, 2019, pp. 4602–4609.
[43] S. Vashishth, S. Sanyal, V. Nitin, and P. Talukdar, "Composition-based multi-relational graph convolutional networks," in *Proc. 8th Int. Conf. Learn. Representations*, 2020.

Zifan Song is currently working toward the Ph.D. degree in computer science and technology with the College of Electronics and Information Engineering, Tongji University, Shanghai, China.

His research interests include deep learning, computer vision, and person search for visual surveillance.

Cairong Zhao received the B.S. degree in electronic information science and technology from Jilin University, Changchun, China, in 2003, the M.S. degree in circuits and systems from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Beijing China, 2006, and the Ph.D. degree in computer science and technology from the Nanjing University of Science and Technology, Nanjing, China, in 2011.

He is currently a Professor with the College of Electronic and Information Engineering, Tongji University, Shanghai, China. He works on visual and intelligent learning, including computer vision, pattern recognition and visual surveillance. He has authored or coauthored over 40 top-rank international conferences and journals in the field, including Conference on Computer Vision and Pattern Recognition, International Conference on Machine Learning, International Conference on Learning Representations, AAAI, Association for Computing Machinery's annual conference on multimedia, International Conference on Computer Vision, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and PR.

Dr. Zhao serves as the Reviewer of more than ten AI-related international journals and conferences, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, Conference on Computer Vision and Pattern Recognition, International Conference on Computer Vision, Conference and Workshop on Neural Information Processing Systems, International Conference on Machine Learning, and AAAI, etc. He serves as the Chairperson of the Computer Vision Special Committee of the Shanghai Computer Society.

Guosheng Hu (Senior Member, IEEE) received the Ph.D. degree in computer vision and deep learning from the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, U.K., in 2015.

He was a Postdoctoral Researcher with the LEAR Team, Inria Grenoble Rhone-Alpes, Montbonnot-Saint-Martin, France, from May 2015 to May 2016. He is currently a Senior Researcher with Oosto, Belfast, U.K. His research interests include deep learning, pattern recognition, and biometrics (mainly face recognition).

Duoqian Miao received the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 1997.

He is currently a Professor and a Ph.D. Tutor with the College of Electronics and Information Engineering, Tongji University, Shanghai, China. His interests include machine learning, data mining, Big Data analysis, granular computing, artificial intelligence, and text image processing.

Dr. Miao was the recipient of the Second Prize at Wu Wenjun AI Science and Technology, in 2018. He serves as an Associate Editor for the *International Journal of Approximate Reasoning* and an Editor of the *Journal of Computer Research and Development* (in Chinese). He serves as the Vice President for the International Rough Set Society, an Executive Manager for the Chinese Association for Artificial Intelligence, the Chair for the CAAI Granular Computing Knowledge Discovery Technical Committee, a Distinguished Member for Chinese Computer Federation, and the Vice President for Shanghai Computer Federation and Shanghai Association for Artificial Intelligence.