Group Re-identification with Group Context Graph Neural Networks

Ji Zhu, Hua Yang, *Member, IEEE,* Weiyao Lin, *Senior Member, IEEE,* Nian Liu, Jia Wang, and Wenjun Zhang, *Fellow, IEEE*

***Abstract*—Group re-identification aims to match groups of people across disjoint cameras. In this task, the contextual information from neighbor individuals can be exploited for re-identifying each individual within the group as well as the entire group. However, compared with single person re-identification, it brings new challenges including group layout and group membership changes. Motivated by the observation that individuals who are close together are more likely to keep in the same group under different cameras than those who are far apart, we propose to model each group as a spatial K-nearest neighbor graph (SKNNG) and design a group context graph neural network (GCGNN) for graph representation learning. Specifically, for each node in the graph, the proposed GCGNN learns an embedding which aggregates the contextual information from neighbor nodes. We design multiple weighting kernels for neighborhood aggregation based on the graph properties including node in-degrees and spatial relationship attributes. We compute the similarity scores between node embeddings of two graphs for group member association and obtain the matching score between the two graphs by summing up the similarity scores of all linked node pairs. Experimental results on three public datasets show that our approach performs favorably against state-of-the-art methods and achieves high efficiency.**

***Index Terms*—group re-identification, spatial K-NN graph, group context graph neural network.**

I. Introduction

PERSON re-identification (re-ID) aims to retrieve a person in non-overlapping camera views. Significant progress has been made on this task and recent methods have achieved promising results on several benchmark datasets [1, 2, 3, 4]. However, in real crowded scenes, people usually walk with others in a group. Person re-ID only focuses on isolated individuals and thus suffers from frequent intra-group occlusions. In contrast, group re-ID aims to match groups of people across disjoint camera views. In this task, the additional contextual information, such as appearances and spatial layouts of neighbors, can be exploited to help re-identify each individual within groups and further understand group activities. Hence, group re-ID has wide applications in video surveillance including multi-person tracking, group retrieval, and crowd analysis.

J. Zhu is with the Department of Electronic Engineering, Shanghai Jiao Tong University and Visbody Inc., Shanghai 200240, China (e-mail: jizhu1023@gmail.com).

H. Yang, W. Lin, J. Wang, and W. Zhang are with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: {hyang, wylin, jiawang, zhangwenjun}@sjtu.edu.cn).

N. Liu is with the Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE. (e-mail: liunian228@gmail.com).

Hua Yang is the corresponding author.

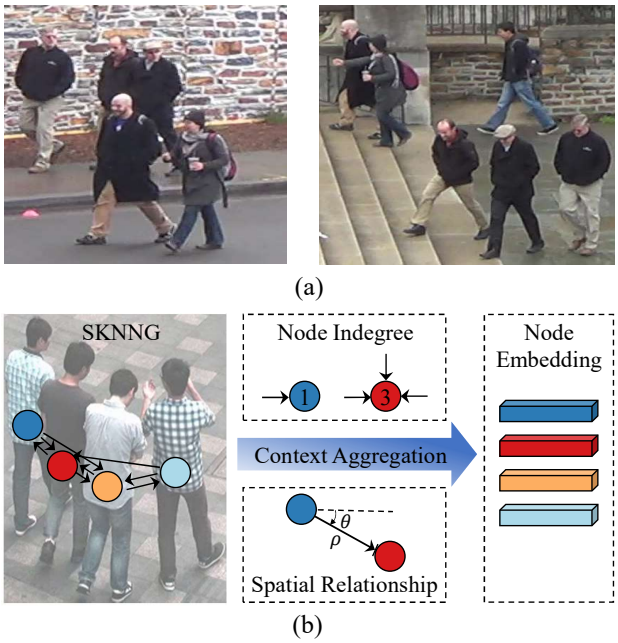


Fig. 1. (a) An example pair of group images undergoing changes of the group layout and membership. (b) Brief illustration of the proposed approach. We model the group as a SKNNG (K=2) where the nodes in different colors represent different individuals in the group. The directed edge from node i to node j indicates that j is one of the K-nearest neighbors of i. Our network applies context aggregation based on both the node indegree and spatial relationship between neighbor nodes to generate node embeddings for graph matching.

However, rich contextual information is a double-edged sword. Despite its benefits, it also brings two new challenges as shown in Fig. 1(a). First, the group layout is highly non-rigid and the relative locations of group members may vary under different cameras. Second, the group size and membership may change over time since people often join or leave the group. Thus, good descriptors of a group should not only take advantage of useful contextual cues but also keep robust to dynamic changes of the group appearance.

Most existing methods [5, 6, 7, 8, 9] consider a group as a whole and extract global or semi-global features which are sensitive to the group layout and membership changes. Different from these methods, Xiao et al. [10, 11] consider a group as a set of multi-grained objects including individuals, two-people subgroups, three-people subgroups, and the entire group. They solve the group matching as a multi-grained object association problem, and thus enhance the robustness to intra-group variances. However, the matching process involves

a full combination of all multi-grained objects with many unstable subgroups, which not only increase the computation complexity but also could degrade the performance. Since the number of group images in the public group re-ID dataset is very limited, Huang [12] et. al. apply a domain adaption method to tranfer images from the person re-ID dataset to the group re-ID dataset style for individual feature learning. They regard a group as an undirected complete graph where each node corresponds to an individual in the group and adopt a graph neural network to learn the group representation. However, the edge between two nodes in the graph only encode the similarity of their appearance features but does not consider the their spatial relationship. Besides, since the graph is complete, the neighborships of individuals are not fully exploited.

In real situations, group members with close proximity to each other are more likely to appear together under different cameras than those who are far apart, and thus are more stable and reliable for group matching. Motivated by this observation, we propose to model a group as a spatial K-nearest neighbor graph (SKNNG) as shown in Fig. 1(b). In this graph, each node i corresponds to an appearance feature vector of a group member. Two nodes i and j are connected by a directed edge from node i to node j if j is one of the K-nearest neighbors of i in spatial space. We associate each edge with spatial relationship attributes including the relative distance and orientation between the two nodes. Besides edge attributes, the node in-degree (i.e. the number of edges pointing to a node) is also useful to characterize the group layout. It partly reflects the importance of each node, since an individual who is the nearest neighbor of most other group members is more likely to be a stable member of this group rather than a person who temporarily walks close to the group and then walks away. Thus, we can rely more on the individual with high node in-degree for group matching.

To learn a graph representation which fully exploits the information encoded in the SKNNG, we design a novel Group Context Graph Neural Network (GCGNN). For each node in the graph, our network generates an embedding which integrates the contextual information from neighbors as shown in Fig. 1(b). More specifically, we define multiple weighting kernels for neighborhood context aggregation based on the node in-degree and spatial relationship (i.e. relative orientation and distance) between neighbor nodes. The features aggregated by different kernels are combined with the original individual feature and mapped to the node embeddings for group association. The weighting kernels are designed to have a certain degree of tolerance for layout misalignment, and the feature combination learned by the network can make a trade-off between the individual appearance and neighborhood contexts which enhances the robustness to group layout and membership changes.

Like Xiao et al. [10, 11], we not only find a match for the entire group but also link the corresponding members between the group pair. When comparing two graphs, we compute a similarity matrix between node embeddings in these two graphs and associate corresponding nodes by solving an assignment problem. The sum of similarity scores for the linked node pairs is regarded as the matching score between the two graphs. Different from [10, 11] where a full combination of all multi-grained objects are considered for group matching, our approach only involves as many node embeddings as group members in the matching objective function, and thus is much more computation-efficient. Besides, since we filter out the unreliable contextual information based on the neighborship, the matching performance are also improved.

The contributions of this work are summarized as follows:

- We propose to model the group as a directed spatial K-NN graph which encodes both the neighborship and the spatial relationship of group members for the first time.
- We design a group context graph neural network with new neighborhood aggregation mechanisms based on the properties of SKNNG including the node indegree and the spatial relationship (i.e. relative distance and orientation) between neighbor nodes.
- Experimental results on three public group re-ID datasets show that our approach achieves high efficiency and performs favorably against state-of-the-art group re-ID methods without using any additional training data from other person re-ID datasets.

II. Related Works

A. Single Person Re-identification

Person re-ID has attracted great attention in both computer vision research and industry communities. In general, existing person re-ID methods either focus on designing appearance representation [13, 14, 15, 16, 17, 18] or learning a distance metric in the feature space [19, 20, 21, 22, 23, 24]. For representation learning, most recent works adopt deep learning based methods. For example, Ahmed et al. [15] propose a deep convolutional network which computes the cross-input neighborhood differences on mid-level features to capture the local relationship between two input person images. They design a patch summary layer to further produce a holistic representation of the neighborhood difference map. Chen et al. [16] integrate CNN and RNN to learn the spatial-temporal fusion features of input sequences for video-based person re-identification. Sun et al. [18] design a network named Part-based Convolutional Baseline (PCB) to learn part-level features and further propose an adaptive part pooling method which refines the within-part consistency. For metric learning, Bak et al. [21] separate the metric into independent texture and color components and propose a one-shot learning algorithm for person re-ID. Hermans et al. [22] propose a new variant of triplet loss named batch hard loss which improves the performance and makes the network easier to converge compared to the traditional triplet loss with offline hard-mining.

Besides exploiting the individual apearance information, several works [25, 26, 27, 28] introduce the group information to improve the performance of single person re-ID in videos. Bialkowski et al. [25] learn to assign each person to a role in a sports team and use the role labels as group context features to aid person re-ID. Ukita et al. [26] group the people based on spatial-temporal features of their trajectories and combine three group features and individual features for

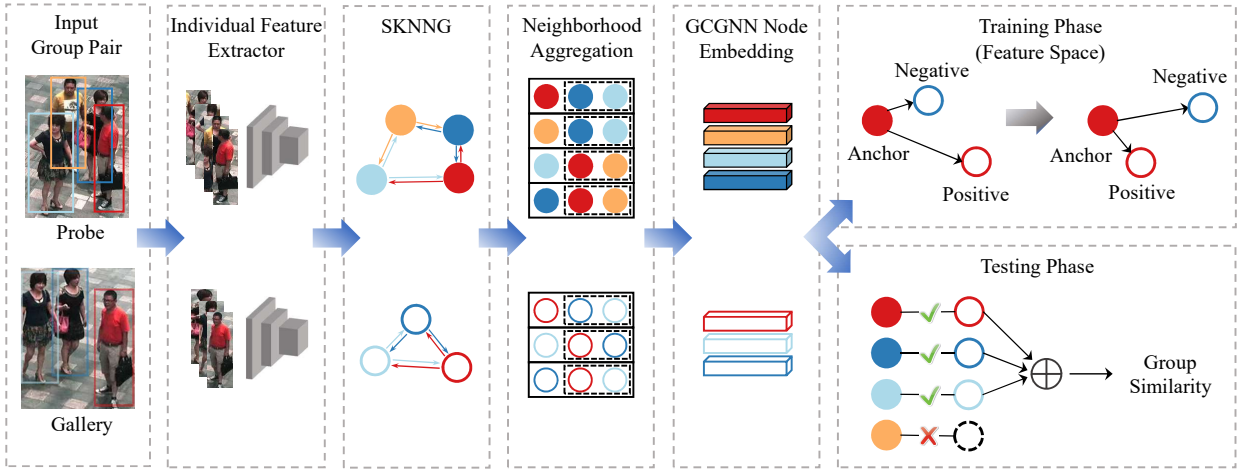


Fig. 2. Proposed pipeline of group re-ID. Given the individual detections in an input group image pair, we extract the appearance feature of each individual and build the SKGNN based on the group layout. We use the proposed GCGNN to apply neighborhood aggregation and generate node embeddings with contextual information. In the training phase, we adopt the triplet loss so that the distance between node embeddings corresponding to the same identity is closer than node embeddings corresponding to different identities in feature space. In the testing phase, we find the matched individual pairs between two group images and use the sum of their similarity scores as the matching score for the input group pair (best viewed in color).

person re-ID. Cao et al. [27] detect the co-traveler set of each person based on their trajectories and propose a pair matching scheme to measure the distance between co-traveler sets. The co-traveler set information is then integrated for individual matching. Assari et al. [28] incorporate multiple Personal, Social and Environmental (PSE) constraints into an energy minimization function to solve person re-ID across cameras in crowded scenes. The PSE constraints include the information of individual appearances, individual speeds, social groups, and the transition probabilities between gates.

B. Group Re-identification

Group re-ID requires to compute the similarities between group images. In this task, both the individual appearance and group layout information can be exploited for group matching. The group-level information is also useful for re-identifying every single person in groups. However, compared with single person re-ID, the task of group re-ID is less studied. Only a few works [5, 6, 7, 8, 9, 10, 11, 12] have been proposed for group re-ID. Most of them [5, 6, 7, 8, 9] regard a group as a whole and extract global or patch features directly on the group image. For example, Wei et al. [5] divide the image into multiple patches by the center rectangular ring based and block based ratio-occurrence descriptors. They propose a top k-match model to associate local patches of two group images. Cai et al. [6] exploit covariance descriptors to represent group images, which capture both the appearance and statistical properties of image patches. Zhu et al. [7] learn the salience channels to filter out unreliable patch matches and propose a consistent matching algorithm for the association of group image pairs. Lisanti et al. [8, 9] first learn a dictionary of sparse atoms on patches extracted from single person images and then obtain sparsity-driven residual representations for group images based on the learned dictionary.

Different from these methods, to enhance the robustness against group layout and membership changes, Xiao et al. [10, 11] consider a group as a set of multi-grained objects (i.e. individuals, two-people subgroups, three-people subgroups, and the entire group) and explicitly associate each group member while applying group matching. This approach uses a sophisticated multi-order matching algorithm to solve the problem of group association with a full combination of all multi-grained objects, and thus has a relatively high computation complexity. Huang [12] et. al. propose to tranfer images from the person re-ID dataset to the group re-ID dataset style to address the lack of training samples in public group re-ID datasets. They regard a group as an undirected graph and adopt a graph neural network to learn the group representation. However, since the graph is complete and the edge between two nodes in the graph only encode the similarity of their appearance features, the spatial relationship (i.e. relative distance and orientaion) and neighborship of individuals are not fully exploited for group matching.

C. Other Group Infomation Based Works

Besides person re-ID, the group information has also been explored to address many other tasks such as visual tracking [29], behavior analysis [30], and vehicle re-ID [31]. Chen et al. [29] propose a elementary grouping graph to model the social grouping behavior of pairwise targets and integrate the group information into the tracklet affinity model to improve data association for multi-target tracking. Alameda et al. [30] propose a dataset for multimodal group behavior analysis. Bai et al. [31] utilize a online clustering method to parition samples within each vehicle ID into a few groups based on the intra-class variance attributes. They generate triplet samples across different vehicle IDs as well as different groups within the same vehicle ID to learn discriminative features for vehicle re-ID.

D. Graph Neural Networks

Graph Neural Networks (GNNs) are introduced in [32, 33], which generalize the neural network to process graph-

structured data. GNNs have been used in various vision tasks such as object detection [34], semantic segmentation [35, 36, 37, 38], and visual question answering [39, 40]. The key idea of GNNs is to generate node embeddings by aggregating neighborhood information using neural networks. Recently, there is an increasing interest in extending convolution operations on graph data for neighborhood aggregation. For example, Kipf et al. [41] propose the Graph Convolutional Networks (GCNs) which aggregate information from 1-step neighborhoods around each node using convolutional operation. Velivckovic et al. [42] propose the Graph Attention Networks (GATs) which adopt a self-attention strategy to compute the weights for neighbor node aggregation.

However, most of these GNN models are applied for undirected graphs and can only exploit one-dimensional edge features encoded in the adjacent matrix which indicates node connectivities or edge weights. Different from these GNN models, our GCGNN considers multi-dimensional edge attributes including the connectivity, edge direction, relative distance, and relative orientation in the neighborhood aggregation process, thus fully exploits the topological context information encoded in the group graph.

III. Proposed Method

Given a probe group image captured in camera A, we aim to compare it with each gallery group image from camera B and find its matches. Fig. 2 shows the pipeline of our method. Given individual detections in a pair of group images, we first extract appearance features of these individuals and build a SKNNG for each group image respectively based on the layout information. Then we exploit the proposed GCGNN to integrate neighborhood contexts into the embedding for each node. Based on the similarity of node embeddings between the two group images, we solve the node assignment problem and use the sum of similarity scores between the associated node pairs as the final matching score of the group image pair. Following the practice of MGR [10, 11], we use the manually annotated ground truth (-GT) or an automatic detection method of [43] (-auto) to get detections of individuals in a group.

A. Graph Structure

To build a graph on a group image, a convenient choice is to use a complete graph so that the connections between all individuals can be exploited. However, as people often join or leave the group under different cameras, there exist many unstable connections in this graph. Taking these connections into consideration will bring interference to group matching. Based on the observation that the connections between individuals who are close together are usually more stable than the connections between those who are far apart, we propose to construct a spatial K-nearest neighbor graph (SKNNG) $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ for each group image. As shown in Fig. 1(b), each node $i \in \mathcal{V}$ is associated with an appearance feature vector of the corresponding individual in the group. Two nodes i and j are connected by a directed edge $e_{ij} \in \mathcal{E}$ from node i to node j, if j is one of the K-nearest neighbors of i in spatial space. Besides neighborship, the edge of SKNNG also encodes the spatial relationship between nodes. With each edge e_{ij}, we associate a vector (ρ_{ij}, θ_{ij}) in polar coordinates from node i to node j to represent the relative distance and orientation between the bounding box centers of the two nodes.

Compared with the undirected complete graph used in [12], the proposed SKNNG has three benefits: (i) It considers the most relevant neighborhood contexts for each node while filtering out the connections which are less reliable for graph matching. (ii) Besides neighborships, the graph edges also encode relative distances and orientations between nodes, which reflect the spatial layout of a group more comprehensively. (iii) The node in-degree (i.e. the number of edges pointing to a node) partly reflects the importance of each node and can be an additional useful metric to characterize the group layout. The intuition behind the last point is that a high-indegree individual who is the K-nearest neighbor of most others usually lies in the central region of the group and is more likely to be a stable member for group matching.

B. Individual Appearance Feature

To perform group re-ID, we first need to extract the appearance feature of each individual in a group. Our framework can be applied independently with any individual feature extraction methods. Following the practice of [11], in this work, we exploit both the handcrafted method and deep convolutional neural network (CNN) for a meaningful comparison.

As shown in Fig. 3, we adopt a siamese framework for feature learning. Given an input pair of individual images (obtained from person detection results and resized to a unified resolution), we first exploit either a conventional handcrafted algorithm or a deep CNN model to extract the intermediate feature for each individual. Specifically, for the handcrafted method, we use the same color and texture features as [10, 11]. We divide each person image into 18 equal-sized blocks along the vertical direction and extract multiple features including HSV, RGB, LAB, YIQ, YCbCr color histograms and Gabor texture features from each block. These features are then concatenated to form an 8,064-dimensional intermediate feature. For the deep CNN model, we apply the ResNet-34 [44] pre-trained on ImageNet [45] as the base network and use the global average pooling layer to generate a 512-dimensional intermediate feature. The intermediate feature is then mapped to a 512-dimensional individual appearance feature by a fully connected layer. Similar to many metric learning approaches, the function of this fully-connected layer is to learn a projection on intermediate features so that the distance between individual appearance features with the same identity is closer than individual appearance features with different identities.

Similar with [46], we apply three losses to learn individual appearance features, including two identification losses and a verification loss. For identification, we apply an identity classifier with cross-entropy loss on the individual appearance feature in each branch respectively. For verification, we first apply the $L2$-normalization on the two individual appearance features and compute the Hadamard product of them as the interacted feature. Then, we add a binary classifier with cross-entropy loss on the interacted feature to predict the similarity

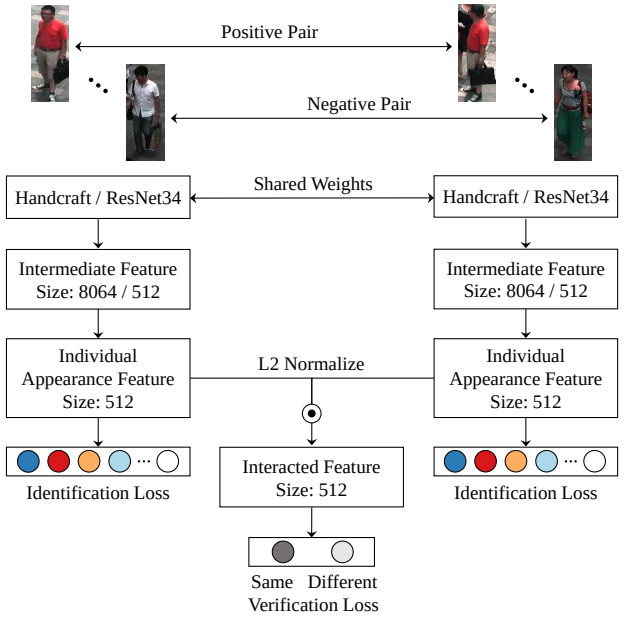


Fig. 3. Individual feature extraction network. The symbol $\odot$ denotes the Hadamard product. We use a siamese architecture with both identification and verification losses for feature learning. Given two individual images as input, each branch first extracts the intermediate feature by either a handcrafted algorithm or a ResNet34 base network. The intermediate feature are then mapped to the individual appearance feature by a fully connected layer.

between the input individual image pair. The identification and verification losses are complementary to each other and improve the discriminative ability of the learned appearance features. In the training phase of deep CNN model, the weights of ResNet34 is fine-tuned on the training data (i.e., jointly updated with the weights of individual appearance feature and interacted feature layers).

C. Group Context Graph Neural Network

After extracting the individual appearance features of a group, we regard these features as node features and input them to a graph neural network (GNN) for graph representation learning. Many GNN variants have been proposed recently and have achieved state-of-the-art results on both node and graph classification tasks. However, popular GNN variants such as GCNs [41] are applied on the adjacent matrix and cannot deal with edges associated with multi-dimensional attributes. To fully exploit information in the SKNNG, we propose a new group context graph neural network (GCGNN) to aggregate context information from local neighbors for each node and generate node embeddings for group matching. As described in Section III-A, besides neighborhood connections, there are two more important group layout properties encoded in the SKNNG. The first one is the node in-degree which represents the proximity of an individual to the other members in the group. The second one is the spatial relationship between neighbor nodes. As explored in [47], the relative positions of people in an image are influenced by their social relationship that is invariant to different cameras. Thus, the spatial relationship between individuals can be a useful cue for group re-identification. We encode the spatial relationship in the edge attribute as a vector (ρ_{ij}, θ_{ij}) in polar coordinates. Different from [10, 11] which extract spatial features independently from appearance features, we exploit the spatial information by designing two aggregator functions based on the node in-degree and spatial relationship to determine the weight of each appearance feature for aggregation. In the following description, we use the subscript d to denote the function or variable for node in-degree based aggregation and use the subscript s to denote the function or variable for spatial relationship based aggregation.

Node In-degree Based Aggregation. Let $\mathbf{h}_i^l$ denote the feature of node i at the l-th layer of GCGNN. Based on the in-degree information, we define a function AGG_d to aggregate the features of neighbors around node i at the $(l-1)$-th layer $\left\{\mathbf{h}_j^{l-1}, \forall j \in \mathcal{N}(i)\right\}$ as:

$$\begin{aligned}\mathbf{h}_{\mathcal{N}(i),d}^l &= \mathrm{AGG}_d\left(\left\{\mathbf{h}_j^{l-1}, \forall j \in \mathcal{N}(i)\right\}\right) \\ &= \mathbf{W}_d^l \sum_{j\in\mathcal{N}(i)} \frac{\exp(\alpha^2 d_j)}{\sum_{j\in\mathcal{N}(i)} \exp(\alpha^2 d_j)} \mathbf{h}_j^{l-1},\end{aligned} \tag{1}$$

where d_j denotes the in-degree of node j and α is a learnable parameter to control the influence of the node in-degree on aggregation weights. The sum of aggregation weights is normalized to 1. The matrix $\mathbf{W}_d^l$ maps the weighted average of neighbor features at the previous layer to the aggregation embedding $\mathbf{h}_{\mathcal{N}(i),d}^l$. By (1), the neighbor node with high degree (i.e., it is the K-nearest neighbor of most others in the group) is considered to be more reliable and gets a higher weight in aggregation.

Spatial Relationship Based Aggregation. The spatial relationship information consists of the relative distance and orientation between nodes. Let the vector (ρ_{ij}, θ_{ij}) denote the relative distance and orientation between node i and node j in polar coordinates as shown in Fig. 1(b). For the relative distance, we define a weighting kernel $w_\rho(i,j)$ to aggregate the feature of each neighbor node $j \in \mathcal{N}(i)$ around node i as:

$$w_\rho(i,j) = \frac{\exp(-\beta^2 \rho_{ij})}{\sum_{j\in\mathcal{N}(i)} \exp(-\beta^2 \rho_{ij})}, \tag{2}$$

where β is a learnable parameter to control the influence of the relative distance on aggregation weights. The sum of $w_\rho(i,j)$ is normalized to 1 so that the matching between group images in disjoint cameras is tolerant to the inter-camera scale difference. As shown in Fig. 4, by (2), the neighbor node $j \in \mathcal{N}(i)$ which is closer to the node i gets a higher weight in aggregation. For the relative orientation, we define multiple weighting kernels $\{w_\theta^n\}$ in the form of:

$$\begin{aligned} w_\theta^n(i,j) &= \exp\left(-\gamma^2 \Delta\theta_{ij}^n\right), \\ \Delta\theta_{ij}^n &= \min\left(|\theta_{ij} - \mu_\theta^n|, 2\pi - |\theta_{ij} - \mu_\theta^n|\right), \\ \mu_\theta^n &= \frac{2\pi n}{N},\ n = 0, 1, \cdots, N-1, \end{aligned} \tag{3}$$

where the mean value μ_θ^n indicates the center direction of each weighting kernel, and γ is a learnable parameter which controls the influence of angle difference $\Delta\theta_{ij}^n$ relative to the center direction μ_θ^n on aggregation weights. As shown

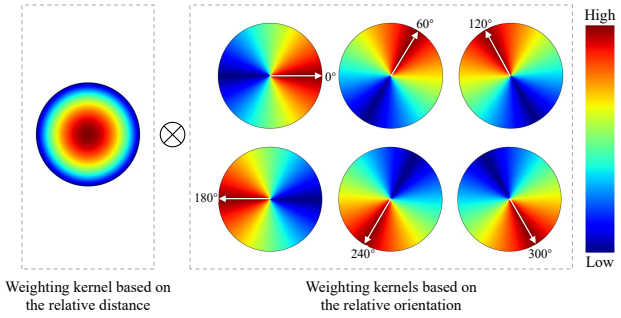


Fig. 4. Weighting kernels based on the spatial relationship between nodes. The left region shows the weighting kernel based on the relative distance. The right region shows 6 weighting kernels with different center directions based on the relative orientation between nodes. The two kinds of weighting kernels are multiplied together to generate the joint weighting kernels for spatial aggregation (best viewed in color).

in Fig. 4, we define $N = 6$ weighting kernels in different directions. For each weighting kernel, the neighbor node $j \in \mathcal{N}(i)$ with the orientation close to the center direction gets a high weight in aggregation. Since the relative distance and orientation should be considered together for spatial relationship, we multiply the weighting kernel based on the distance with each weighting kernel based on the orientation as the joint weighting kernel $w_s^n(i,j) = w_\rho(i,j) \cdot w_\theta^n(i,j)$. We define a function AGG_s to concatenate the aggregated features from all joint weighting kernels and generate the aggregation embedding as:

$$\begin{aligned}\mathbf{h}_{\mathcal{N}(i),s}^l &= \mathrm{AGG}_s\left(\mathbf{h}_j^{l-1}, \forall j \in \mathcal{N}(i)\right) \\ &= \Big\|_{n=1}^{N} \mathbf{W}_{s,n}^l \left(\frac{1}{Z_i^n} \sum_{j \in \mathcal{N}(i)} w_\rho(i,j) w_\theta^n(i,j) \mathbf{h}_j^{l-1}\right),\end{aligned} \tag{4}$$

where $Z_i^n = \sum_{j \in \mathcal{N}(i)} w_\rho(i,j) w_\theta^n(i,j)$ normalizes the sum of neighor weights from the n-th joint weighting kernel to 1. The learnable matrix $\mathbf{W}_{s,n}^l$ maps the weighted average of neighbor node features from each weighting kernel to a compact embedding. The operator $\|_{n=1}^{N}$ denotes the concatenation of features aggregated with N joint weight kernels.

Finally, we apply a linear mapping on the embedding of node i at $(l-1)$-th layer and concatenate it with the neighborhood aggregation embeddings based on the node in-degree as well as the spatial relationship to generate the node embedding at l-th layer as:

$$\mathbf{h}_i^l = \sigma\left(\mathbf{h}_{\mathcal{N}(i),d}^l \,\|\, \mathbf{h}_{\mathcal{N}(i),s}^l \,\|\, \mathbf{W}_o^l \mathbf{h}_i^{l-1}\right), \tag{5}$$

where $\mathbf{W}_o^l$ is a learnable matrix for feature mapping and $\|$ denotes the concatenation operation. Note that $\mathbf{h}_i^0$ is the original individual appearance feature of node i. By (5), both the appearance and spatial layout information are integrated for group matching.

D. Training Strategy

As the number of group images in public group re-ID datasets is very limited, the neural network is prone to overfit the training data. To avoid the overfitting problem, we exploit a two-step training strategy. We first utilize the person detections and identity information provided in the group image dataset to generate person image pairs for the individual network training. Then we fix the weights of the individual network and use it to extract individual appearance features. The proposed GCGNN takes the individual appearance features in a group as input and generates node embeddings based on the group layout context. Each node embedding encodes the appearance and spatial layout information of a subgroup containing an individual of interest and his/her neighbors. The goal of the GCGNN training is to make the distance between the node embeddings corresponding to the same identity smaller than the distance between those corresponding to different identities in the feature space. To achieve this goal, we use the cosine similarity to measure the embedding distance and adopt the triplet loss as:

$$L = \sum_{q}^{Q} \max\left(0, \cos(\mathbf{x}_q^a, \mathbf{x}_q^n) - \cos(\mathbf{x}_q^a, \mathbf{x}_q^p) + m\right), \tag{6}$$

where $\mathbf{x}_q^a$, $\mathbf{x}_q^p$, and $\mathbf{x}_q^n$ denote the $L2$-normalized anchor, positive, and negative embeddings of the q-th node triplet respectively and Q denotes the number of triplets for training. The hyperparameter m is an enforced margin distance between positive and negative node embedding pairs.

Since there are much more negative node pairs than positive node pairs, simply generating all possible node triplets would result in data imbalance and degrade the discriminability of the model. To alleviate the data imbalance problem, we adopt an online hard negative mining strategy. Specifically, in a training mini-batch, we first generate a number of group image triplets. Each group image triplet consists of a probe group image I^a (anchor), a gallery group image I^p (positive) which is the match of I^a, and a randomly sampled gallery group image I^n (negative) which contains a different group from I_a. Based on the SKNNG structure, we build the graph $\mathcal{G}^a$, $\mathcal{G}^p$, and $\mathcal{G}^n$ for the group image I^a, I^p, and I^n respectively. Then we successively select each node in $\mathcal{G}_a$ as an anchor node, consider the matched individual in $\mathcal{G}_p$ as the positive node, and regard other nodes in $\mathcal{G}_p$ and $\mathcal{G}_n$ as negative nodes. Among these negative nodes, we select hard negative nodes whose embeddings generated by the GCGNN are the top T nearest to the anchor node embedding in the feature space. Using the positive node and the selected top T hard negative nodes, we build T node triplets for each anchor node. These node triplets are finally used to train the GCGNN.

E. Group Matching

Given a probe group image and a gallery group image, we use the GCGNN to generate a probe node embedding set and a gallery node embedding set. Each node embedding integrates the information of a subgroup containing an corresponding individual and his/her neighbors. Our goal is to simultaneously associate corresponding probe and gallery node pairs and predict the matching score of the two groups. The matching of the node embeddings can be regarded as subgroup-level

Fig. 5. Example group images in the (a) i-LIDS Group, (b) Road Group, and (c) DukeMTMC Group datasets.

matching. Specifically, we exploit the Hungarian [48] algorithm to make assignments between probe and gallery node embeddings based on their cosine similarity scores. Note that all the node embeddings are L2-normalized before calculating the cosine similarity scores. If the number of individuals in the probe group image is more than that in the gallery group image, there will be unassigned individual(s) in the probe group image, and vice versa. We use the sum of similarity scores between the matched probe-gallery node pairs as the matching score of the two group images, and the gallery group image with the highest matching score is regarded as the match for the probe image.

IV. Experimental Results

A. Datasets and Experimental Settings

We evaluate the proposed group re-ID method on three public datasets: (1) the i-LIDS Group dataset [5] containing 274 images for 64 groups extracted from the original i-LIDS MCTS dataset captured at an airport arrival hall using a multi-camera CCTV network [49], (2) the Road Group dataset [10, 11] with 162 pairs of group images captured from two disjoint cameras, and (3) the DukeMTMC Group dataset [10, 11] with 177 pairs of group images extracted from the original DukeMTMC dataset [50] taken by 8 cameras. Fig. 5 shows some example images of these three datasets. The groups in the i-LIDS Group dataset are more compact but their images suffer from low quality and large illumination changes. The Road Group and DukeMTMC Group datasets have larger groups and thus undergo more group layout and membership variation.

Following the practice of [7, 10, 11], we randomly split each dataset by half into training and test sets for 5 times and use the average Cumulated Matching Characteristic (CMC) score as the evaluation metric. The number of neighbor nodes for aggregation is set to $K = 2$ and the number of orientation

TABLE I
Performances of different aggregation methods on the Road Group dataset.

	Method	R-1	R-5	R-10	R-15	R-20
Handcrafted	W/OA	65.7	85.2	90.1	93.3	95.1
	EA	74.1	90.4	93.8	95.1	96.5
	DA	75.3	92.8	93.6	95.1	97.3
	SA	77.5	92.6	95.1	97.5	97.5
	DA+SA (ours-GT)	**78.8**	**93.8**	**96.5**	**97.8**	**98.5**
Deep Conv	W/OA	73.1	89.6	93.8	95.1	95.6
	EA	80.2	91.4	92.6	96.0	96.3
	DA	81.5	92.1	93.8	95.1	97.5
	SA	82.7	93.8	95.1	96.3	98.3
	DA+SA (ours-GT)	**84.2**	**95.8**	**97.3**	**97.5**	**98.5**

weighting kernels $\{w_\theta^n\}$ is set to $N = 6$. The learnable matrices $\mathbf{W}_d^l$ and $\mathbf{W}_{s,n}^l$ map the input feature to a 128-dimensional embeddding while $\mathbf{W}_o^l$ maps the input feature to a 256-dimensional embedding. Since the number of group images in each group re-ID dataset is very limited, it is prone for the network to overfit the training set. We find that using only one layer for neighborhood context aggregation in the GCGNN alleviates this problem and achieve the best performance. For hard negative mining, we choose $T = 5$ in our experiments. All networks are trained using the Adam solver [51] with a learning rate of 0.01.

B. Ablation Studies

We present extensive ablation studies to demonstrate the effectiveness of our proposed approach on the Road Group dataset. Following the practice of MGR [10, 11], in ablation studies, we use the manually annotated ground truth (-GT) to get detections of individuals in a group.

1) Contribution of each aggregation method: To demonstrate the contribution of the proposed neighborhood aggregation methods adopted in the GCGNN, we compare five methods using different aggregation strategies. Each method is described as follows:

"W/OA": We only use the individual appearance features without neighborhood aggregation for group matching.

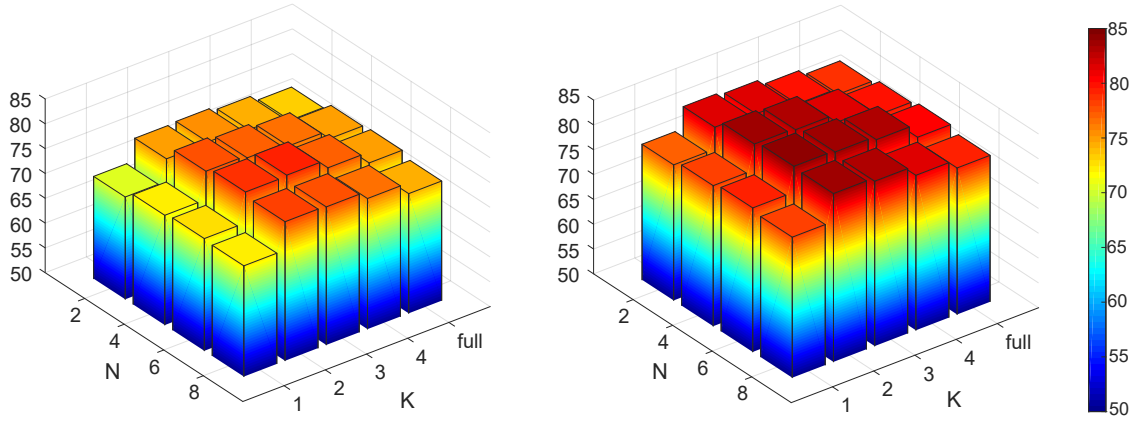
"EA": We simply assign equal weights to all neighbor nodes for the neighborhood aggregation.

"DA": We apply the proposed node in-degree based neighborhood aggregation.

"SA": We apply the proposed spatial relationship based neighborhood aggregation.

"DA+SA": We apply the proposed GCGNN with both node in-degree based and spatial relationship based neighborhood aggregation.

Table I shows the CMC results of each method on the Road Group dataset. The upper part shows the performances based on the handcrafted individual appearance features and the lower part presents the results based on the deep convolutional individual appearance features. When using the same aggregation strategy, the performances based on the deep convolutional features are better than those based on the handcrafted features, which demonstrates the advantage of features extracted by the deep CNN. As we can see from



(a) Handcrafted (b) Deep Conv

Fig. 6. Rank-1 CMC scores on the Road Group dataset with different values of K (number of neighbors in the SKGNN) and N (number of orientation weighting kernels for spatial relationship based aggregation) using (a) handcrafted and (b) deep convolutional individual appearance features as input of the CGCNN, respectively. The label "full" means that we build a complete graph where all the other individuals are considered as neighbors of one individual in the graph. Each bar is colored based on the corresponding rank-1 CMC score (best viewed in color).

the results, both of the two proposed neighborhood aggregation methods make contributions to the performance. The "W/OA" method achieves the poorest results because it only considers the individual appearance without exploiting any neighborhood context information. Both the "DA" and "SA" methods have better performance than the "EA" method because the two proposed aggregation methods integrate more group layout information. Besides the individual appearances and neighborships information, the "DA" method further considers the proximity of each individual to the other group members and the "SA" method encodes the relative distances and orientations between neighbor nodes. The proposed GCGNN with both the "DA" and "SA" achieves the best performance, which demonstrates the complementarity between the "DA" and "SA" methods.

2) Effects of hyperparameters: We also evaluate the effects of hyperparameters including the number K of neighbors in the SKNNG, the number N of orientation weighting kernels $\{w_\theta^n\}$ for spatial relationship based aggregation, and the number T of hard negative samples for hard negative mining.

Fig. 6 shows the rank-1 CMC scores on the Road Group dataset with different values of K and N. For the number K of neighbors in the SKNNG, using only one neighbor performs the worst since the nearest neighbor of an individual is relatively easy to switch. Aggregating context information from 2 or 3 neighbors achieves better performance. Considering more than 3 neighbors involves more unreliable context information for group matching and thus leads to performance degradation. These results also demonstrate the advantage of the proposed SKNNG compared with the complete graph ("full" in Fig. 6). For the number N of orientation weighting kernels $\{w_\theta^n\}$, using few (e.g. 2 or 4) weighting kernels is not able to fully capture the spatial orientation relationships between individuals and thus leads to the suboptimal performance. However, because the viewpoint and group layout may change in different cameras, there maybe some changes of relative orientation between the same individual pair in two group images.

TABLE II
RANK-1 CMC SCORES ON THE ROAD GROUP DATASET WITH DIFFERENT VALUES OF T. THE LABEL "ALL" MEANS THAT WE EXPLOIT ALL NEGATIVE NODES IN EACH BATCH FOR NETWORK TRAINING.

T	3	5	7	9	all
Handcrafted	76.5	**78.8**	78.5	78.0	77.8
Deep Conv	82.2	**84.2**	84.0	83.7	83.5

Thus, narrowing the orientation bins and defining more (e.g. 8) weighting kernels do not necessarily improve the performance as shown in Fig. 6. In our experiments, using 6 orientation weights kernels achieves the best performance, which is fine-grained enough to capture the spatial characteristics while maintaining a certain tolerance to the changes of viewpoint and group layout.

Table II shows the rank-1 CMC scores on the Road Group dataset with different values of T. Setting T to 5 achieves the best performance. Exploiting all negative nodes in each batch without hard negative mining degrades the performance since the model training may be overwhelmed by easy negatives. Using only the top 3 hardest negative nodes to build triplets for training also reduces the performance.

3) Comparison with other GNN structures: We further compare our proposed GCGNN against other two typical network structures. First, when comparing a pair of group images, a straightforward alternative solution is to apply a graph pooling layer [12] on the learned local node features to get a fixed-sized graph representation and add a binary classifier to estimate the similarity between these two global-level representations. However, it is difficult for the global-level representation to capture the local appearance and structural information, which leads to a relatively poor matching performance as shown in Table III. In contrast, by directly comparing the node features, we retain the local contextual information and establish the finegrained correspondences between two graphs.

TABLE III
PERFORMANCES OF DIFFERENT GNN STRUCTURES ON THE ROAD GROUP DATASET.

	Method	R-1	R-5	R-10	R-15	R-20
Handcrafted	GraphPool	17.5	35.1	56.3	62.2	65.2
	GATs	65.2	80.5	86.4	90.6	92.8
	GCNs	68.1	86.4	91.4	93.8	95.3
	GCGNN	**78.8**	**93.8**	**96.5**	**97.8**	**98.5**
Deep Conv	GraphPool	25.7	44.2	67.4	71.6	76.5
	GATs	71.6	87.2	91.1	93.3	93.6
	GCNs	74.1	90.1	93.8	95.1	95.3
	GCGNN	**84.2**	**95.8**	**97.3**	**97.5**	**98.5**

Second, we compare our proposed GCGNN with the popular GCNs using the method in [41] for neighborhood aggregation. As shown in Table III, our GCGNN performs better than the GCNs. It is because the GCNs only encode the connection attributes of graph edges (corresponding to neighborship) and ignore other spatial attributes of edges. In contrast, the proposed GCGNN exploits the node in-degree information in the directed SKNNG and integrates multi-dimensional edge attributes (i.e. relative distance and orientation) into the node embeddings, which minimize the loss of spatial layout information from the input graph.

Third, we compare our proposed GCGNN with the GATs [42]. Similar to our network, GATs also apply neighborhood aggregation based on the learned weights that indicate the importance of each neighbor node. However, the aggregation weights in GATs are generated by a self-attention strategy and thus depend only on the appearance features of neighbor nodes without considering the spatial context information. As shown in Table III, our GCGNN achieves better performance than GATs.

4) Effects of jointly using identification and verification losses: As described in Section III-B, our individual feature extraction network is trained with both identification and verification losses. To demonstrate the benefit of jointly using these two kinds of losses together, we train the individual feature extraction network with just identification loss, just verification loss, and both identification and verification losses, respectively. Then we apply the learned individual features in group re-ID to evaluate their performances on the Road Group dataset under the setting of "W/OA" (i.e., only using the individual appearance features without neighborhood aggregation for group matching). As shown in Table IV, for both handcrafted and deep convolutional individual features, the features jointly trained with identification and verification losses outperform those features trained with only the identification or verification loss, which demonstrate the complementarity of these two kinds of losses.

5) Group Split and Person Swap: To demonstrate the robustness of our approach to group split and person swap, we select the probe-gallery group pairs undergoing these two kinds of changes from the test set of Road Group dataset and build the "hard set with group split" and "hard set with person swap", respectively. Specifically, for the hard set with group split, we pick out groups in which one or more person leave the group in the probe image or gallery image. For the hard set with person swap, we select the groups in which there is a swap of single identities. Table V shows the group re-ID performances on the two selected hard sets. Note that in this experiment, while the probe images are selected from the two hard sets respectively, the size of the gallery image set for group matching keeps the same with the full test set.

TABLE IV
GROUP RE-ID PERFORMANCES ON THE ROAD GROUP DATASET ("W/OA") USING INDIVIDUAL FEATURES LEARNED WITH JUST IDENTIFICATION LOSS ("I"), JUST VERIFICATION LOSS ("V"), AND BOTH IDENTIFICATION AND VERIFICATION LOSSES ("I+V"), RESPECTIVELY.

	Loss	Rank-1 Score
Handcrafted (W/OA)	I	63.4
	V	62.2
	I+V	**65.7**
Deep Conv (W/OA)	I	70.4
	V	68.1
	I+V	**73.1**

TABLE V
PERFORMANCES ON THE HARD SET WITH GROUP SPLIT ("GS") AND HARD SET WITH PERSON SWAP ("PS"), RESPECTIVELY.

	Hard Set	R-1	R-5	R-10
Handcrafted	GS	73.9	87.0	91.3
	PS	64.7	76.4	76.4
Deep Conv	GS	82.6	91.3	91.3
	PS	70.5	76.4	82.4

On the hard set with group split, it only shows a little effects on the group re-ID performance compared with our results on the full test set (rank-1 score 73.9 VS 78.8 for handcrafted features, 82.6 VS 84.2 for deep convolutional features as shown in Table I). This is because if there is no match obtained through the application of the Hungarian algorithm for a particular graph node, then it denotes the fact that the corresponding individual has left the group. In that case, the individual who leaves the group is not included in the summation of matching scores as shown in Fig. 2 and we will rely on other individuals for group matching.

On the hard set with person swap, we still achieve a competitive performance compared to the results of existing patch-based methods [5, 6, 7, 8, 9] evaluated on the full test set (rank-1 score 70.5 VS 58.6 as shown in Table VII). Since our node embeddings combine the individual appearance features and neighborhood context features together, the individual appearance and spatial layout information can compensate for the variations of each other. The feature combination learned by our network makes a trade-off between the individual appearance and spatial layout information, which enhances the robustness to person swap.

C. Results for Single Person Re-ID

Our group re-ID algorithm is also beneficial to the single person re-ID task. First, the group matching information can narrow down the search scope to the persons in the matched group and reduce the ambiguity in person re-ID. Second, since

TABLE VI
RESULTS OF METHODS USING DIFFERENT MATCHING STRATEGIES FOR SINGLE PERSON RE-ID ON THE ROAD GROUP DATASET.

	Method	Rank-1 Score
Handcrafted	W/OC+W/OG	27.9
	W/OC+WG	69.4
	WC+WG (ours-auto)	**73.6**
	MGR-auto [10, 11]	71.4
Deep Conv	W/OC+W/OG	33.6
	W/OC+WG	72.3
	WC+WG (ours-auto)	**78.7**
	MGR-auto [11]	73.5

our approach not only performs group matching but also learns node embeddings to apply individual association between two group images, the learned node embeddings with contextual information can be directly used as the representation of corresponding persons for person re-ID. To evaluate the effects of our proposed approach on the single person re-ID task, we compare three methods with different matching strategies described as follows:

"W/OC+W/OG": We use the individual appearance features (described in Section III-B) without context information as person representations (W/OC). Given a probe person image, we find the matched person from all persons in the gallery group images based on the cosine similarities of person representations without using any group matching information (W/OG).

"W/OC+WG": We use the individual appearance features (described in Section III-B) without contextual information as person representations (W/OC). Given a probe person image, we first apply our group re-ID algorithm to retrieve its matched group in the gallery and then find the person with the highest cosine similarity score in the matched gallery group image as the matched person (WG).

"WC+WG": We use the node embeddings extracted by the proposed GCGNN with contextual information as person representations (WC). Given a probe person image, we first apply our group re-ID algorithm to retrieve its matched group in the gallery and then find the person with the highest cosine similarity score in the matched gallery group image as the matched person (WG).

Table VI shows the rank-1 CMC score of each method on the Road Group dataset using handcrafted features and deep convolutional features respectively. Following the practice of MGR [10, 11], we use pedestrian detections automatically generated by a pedestrian detector [43] to identify individuals in a group image (-auto). As we can see, the "W/OC+W/OG" method gets the poorest performance. Since this method searches the matched individual among all individuals in the entire galley set without using the group matching information, it is prone to get a matched person in other group with a similar appearance as shown in Fig. 7(a). When both applying group matching to narrow down the person search scope, the method using the proposed GCGNN node embeddings ("WC+WG") achieves higher accuracy than the method using the original individual features without integrating the contextual information ("W/OC+WG"). This is because the GCGNN node

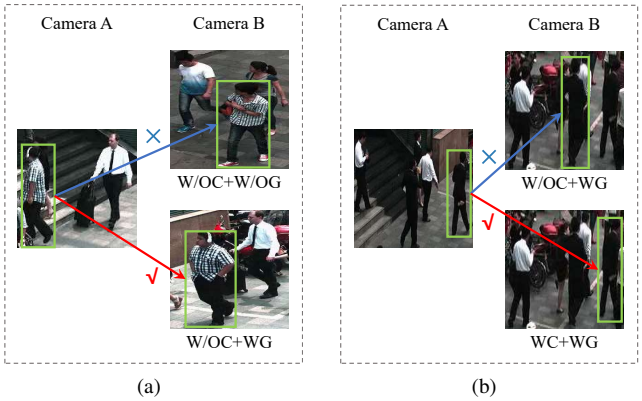


Fig. 7. Examples of single re-ID results using different matching strategies. (a) The "W/OC+W/OG" gets the incorrect match (blue arrow) while the "W/OC+WG" gets the correct match (red arrow). (b) The "W/OC+WG" gets the incorrect match while the "WC+WG" gets the correct match (best viewed in color).

embeddings provide additional cues to reduce the ambiguity and avoid matching to the group member with a similar appearance as illustrated in Fig. 7(b). Both of our methods using handcrafted features and deep convolutional features perform better than the MGR [10, 11], which demonstrates the effectiveness of our proposed group re-ID algorithm for improving the single person re-ID performance.

D. Comparison with State-of-the-art Methods

Following the practice of MGR [10, 11], in this experiment, we use bounding boxes predicted by an automatic pedestrian detection method [43] to identify individuals in groups (-auto). Table VII shows our performance on the i-LIDS Group dataset, Road Group dataset, and DukeMTMC Group dataset against state-of-the-art group re-ID methods including the CRRO-BRO [5], Covariance [6], PREF [8, 9], BSC+CM [7], DoTGNN [12], and MGR [10, 11]. Since the MGR [10, 11] presents their results using handcrafted features (hand) and deep convolutional features (conv) respectively, to demonstrate the effect of our proposed GCGNN, we also evaluate our method using handcrafted features (hand) and deep convolutional features (conv) respectively. As described in Section III-B, for handcrafted features, we use the same color and texture features as the MGR (hand) [10, 11] except that we further project the original appearance features to the 512-dimensional features through a fully conntected layer. For deep convolutional features, we also exploit the same base network (i.e. ResNet-34) as the MGR (deep) [11]. Unlike the MGR [10, 11] which has a separated branch to explicitly extract spatial features, our GCGNN integrate both appearance and spatial characteristics together into node embeddings. Note that we compute the similarity between two group images by directly comparing a probe image with a given gallery image while the MGR [10, 11] exploits additional information from the entire gallery group image set to adjust the similarity score. Even though, both the handcrafted feature version and the deep convolutional feature version of our approach perform better than the MGR [10, 11] as shown in Table VII. Our method

TABLE VII
COMPARISON WITH STATE-OF-THE-ART METHODS FOR GROUP RE-ID ON THREE DATASETS.

Method	i-LIDS Group					Road Group					DukeMTMC Group				
	R-1	R-5	R-10	R-15	R-20	R-1	R-5	R-10	R-15	R-20	R-1	R-5	R-10	R-15	R-20
CRRRO-BRO [5]	23.3	54.0	69.8	76.7	82.7	17.8	34.6	48.1	57.5	62.2	9.9	26.1	40.2	54.2	64.9
Covariance [6]	26.5	52.5	66.0	80.0	90.9	38.0	61.0	73.1	79.0	82.5	21.3	43.6	60.4	70.3	78.2
PREF [8, 9]	30.6	55.3	67.0	82.0	92.6	43.0	68.7	77.9	82.2	85.2	22.3	44.3	58.5	67.4	74.4
BSC+CM [7]	32.0	59.1	72.3	82.4	93.1	58.6	80.6	87.4	90.4	92.1	23.1	44.3	56.4	64.3	70.4
DoTGNN [12]	-	-	-	-	-	74.1	90.1	92.6	-	**98.8**	53.4	72.7	80.7	-	88.6
MGR-auto (hand) [10, 11]	37.9	64.5	79.4	91.5	93.8	72.3	90.6	94.1	97.1	97.5	47.4	68.1	77.3	83.6	87.4
MGR-auto (conv) [11]	38.8	65.7	82.5	93.8	**98.8**	80.2	93.8	96.3	97.5	97.5	48.4	75.2	89.9	93.3	94.4
Ours-auto (hand)	39.4	66.3	83.7	92.5	95.0	75.1	92.1	95.6	97.3	97.3	49.8	70.5	81.6	87.0	90.7
Ours-auto (conv)	**41.9**	**68.1**	**86.9**	**94.4**	98.1	**81.7**	**94.3**	**96.5**	**97.5**	97.8	**53.6**	**77.0**	**91.4**	**93.6**	**94.8**

TABLE VIII
RUNNING TIME ON THE ROAD GROUP AND DUKEMTMC GROUP DATASETS.

	Method	Road Group	DukeMTMC Group
All image pairs	MGR	11.5 min	18.9 min
	Ours	**1.7 sec**	**2.0 sec**
Per image pair	MGR	1.0e-1 sec	1.4e-1 sec
	Ours	**3.6e-3 sec**	**3.7e-3 sec**

also performs favorably against the DoTGNN [12] without using any augmented training samples from other person re-ID datasets. These results demonstrate the effectiveness of the proposed approach.

E. Running Time Analysis

We compare the running time of our method with that of the MGR [10, 11] which achieves the best performance among existing methods without using additional training data from person re-ID datasets. Like the MGR [10, 11], we conduct experiments on a platform with an 8-core i7-7700@3.6GHz CPU and an NVIDIA GeForce GTX 1080 GPU. In Table VIII, we present (i) the running time of the entire matching process (i.e. comparing all the image pairs in the test set), and (ii) the average running time for computing the similarity of a single group image pair in the Road Group and DukeMTMC Group datasets. Note that we follow the practice of the MGR [10, 11] and exclude the time consumed for object detection and individual feature extraction. As shown in Table VIII, our approach is much more efficient. This is because in the MGR [10, 11], to overcome the challenges of group layout and membership changes, they involve a full combination of all multi-grained objects (i.e. individuals, two-people subgroups, three-people subgroups, and the entire group) and apply a sophisticated multi-order matching algorithm for group association. Different from that, we only consider the subgroup contextual information from K-nearest neighbors with higher reliability and can directly apply the Hungarian algorithm for both individual and group matching.

V. CONCLUSION

In this work, we propose a group re-ID approach which not only takes advantage of the contextual information in group images but also is robust to dynamic changes of the group layout and membership. We model each group as a spatial K-NN graph and design a group context graph neural network for graph representation learning. We define two new neighborhood aggregation mechanisms based on graph properties including the node in-degree and spatial relationship between individuals. Experimental results on three public group re-ID datasets demonstrate the effectiveness and efficiency of the proposed approach against state-of-the-art methods.

ACKNOWLEDGMENT

This work was supported in part by National Natural Science Foundation of China (NSFC, Grant No. 61771303, 61771305, 61971277), Science and Technology Commission of Shanghai Municipality (STCSM, Grant Nos. 19DZ1209303, 18DZ1200102, 18DZ2270700), Shaanxi Province Technological Innovation Guidance Special Fund (2020QFY01-04), and SJTUYitu/Thinkforce Joint Laboratory for Visual Computing and Application.

REFERENCES

[1] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Proceedings of the Asia Conference on Computer Vision*, 2012.

[2] W. Li and X. Wang, "Locally aligned feature transforms across views," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[3] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[4] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1116–1124.

[5] Z. Wei-Shi, G. Shaogang, and X. Tao, "Associating groups of people," in *Proceedings of the British Machine Vision Conference*, 2009, pp. 23–1.

[6] Y. Cai, V. Takala, and M. Pietikainen, "Matching groups of people by covariance descriptor," in *Proceedings of the International Conference on Pattern Recognition*, 2010, pp. 2744–2747.

[7] F. Zhu, Q. Chu, and N. Yu, "Consistent matching based on boosted salience channels for group re-identification," in *Proceedings of the IEEE International Conference on Image Processing*, 2016, pp. 4279–4283.

[8] G. Lisanti, N. Martinel, A. Del Bimbo, and G. L. Foresti, "Group re-identification via unsupervised transfer of sparse features encoding," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2449–2458.

[9] G. Lisanti, N. Martinel, C. Micheloni, A. Del Bimbo, and G. L. Foresti, "From person to group re-identification via unsupervised transfer of sparse features," *Image and Vision Computing*, vol. 83-84, pp. 29–38, 2019.

[10] H. Xiao, W. Lin, B. Sheng, K. Lu, J. Yan, J. Wang, E. Ding, Y. Zhang, and H. Xiong, "Group re-identification: Leveraging and integrating multi-grain information," in *Proceedings of the ACM International conference on Multimedia*, 2018, pp. 192–200.

[11] W. Lin, Y. Li, H. Xiao, S. John, J. Zou, H. Xiong, J. Wang, and M. Tao, "Group re-identification with multi-grained matching and integration," *IEEE Transaction on Cybernetics*, 2019.

[12] Z. Huang, Z. Wang, W. Hu, C.-W. Lin, and S. Satoh, "Dot-gnn: Domain-transferred graph neural network for group re-identification," in *Proceedings of the ACM International conference on Multimedia*, 2019, pp. 1888–1896.

[13] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Proceedings of the International Conference on Pattern Recognition*, 2014, pp. 34–39.

[14] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 152–159.

[15] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3908–3916.

[16] L. Chen, H. Yang, J. Zhu, Q. Zhou, S. Wu, and Z. Gao, "Deep spatial-temporal fusion network for video-based person reidentification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2017.

[17] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3960–3969.

[18] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 480–496.

[19] S. Paisitkriangkrai, C. Shen, and A. Van Den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1846–1855.

[20] S. Bak and P. Carr, "Person re-identification using deformable patch metric learning," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2016, pp. 1–9.

[21] S. Bak and P. Carr, "One-shot metric learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2990–2999.

[22] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[23] Z. Liu, D. Wang, and H. Lu, "Stepwise metric promotion for unsupervised video person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2429–2438.

[24] H.-X. Yu, A. Wu, and W.-S. Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 994–1002.

[25] A. Bialkowski, P. Lucey, X. Wei, and S. Sridharan, "Person re-identification using group information," in *International Conference on Digital Image Computing: Techniques and Applications*, 2013.

[26] N. Ukita, Y. Moriguchi, and N. Hagita, "People re-identification across non-overlapping cameras using group features," *Computer Vision and Image Understanding*, vol. 144, pp. 228–236, 2016.

[27] M. Cao, C. Chen, X. Hu, and S. Peng, "From groups to co-traveler sets: Pair matching based person re-identification framework," in *Proceedings of the IEEE International Conference on Computer Vision Workshop*, 2017.

[28] M. S. Assari, H. Idrees, and M. Shah, "Human re-identification in crowd videos using personal, social and environmental constraints," in *Proceedings of the European Conference on Computer Vision*, 2016.

[29] X. Chen, Z. Qin, L. An, and B. Bhanu, "An online learned elementary grouping model for multi-target tracking," in *Proceedings of the International Conference on Pattern Recognition*, 2014.

[30] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe, "Salsa: A novel dataset for multimodal group behavior analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1707–1720, 2016.

[31] Y. Bai, L. Yihang, G. Feng, S. Wang, Y. Wu, and L.-Y. Duan, "Group-sensitive triplet embedding for vehicle reidentification," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2385–2399, 2018.

[32] M. Gori, G. Monfardini, and F. Scarselli, "A new model for learning in graph domains," in *IEEE International Joint Conference on Neural Networks*, 2005, pp. 729–734.

[33] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2008.

[34] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation net-

works for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3588–3597.

[35] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with graph lstm," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 125–143.

[36] L. Landrieu and M. Simonovsky, "Large-scale point cloud semantic segmentation with superpoint graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4558–4567.

[37] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *arXiv preprint arXiv:1801.07829*, 2018.

[38] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun, "3d graph neural networks for rgbd semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5199–5208.

[39] D. Teney, L. Liu, and A. van den Hengel, "Graph-structured representations for visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1–9.

[40] M. Narasimhan, S. Lazebnik, and A. Schwing, "Out of the box: Reasoning with graph convolution nets for factual visual question answering," in *Advances in Neural Information Processing Systems*, 2018, pp. 2654–2665.

[41] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proceedings of the International Conference on Learning Representations*, 2017.

[42] P. Velivckovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *Proceedings of the International Conference on Learning Representations*, 2018.

[43] L. Liu, W. Lin, L. Wu, Y. Yu, and M. Y. Yang, "Unsupervised deep domain adaptation for pedestrian detection," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 676–691.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[45] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[46] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person reidentification," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, no. 1, p. 13, 2018.

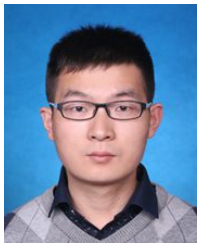
[47] G. Wang, A. Gallagher, J. Luo, and D. Forsyth, "Seeing people in social context: Recognizing people and social relationships," in *Proceedings of the European Conference on Computer Vision*, 2010.

[48] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the society for industrial and applied mathematics*, vol. 5, no. 1, pp. 32–38, 1957.

[49] UK Home Office, "i-lids multiple camera tracking scenario definition," 2008.

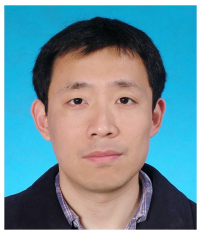
[50] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proceedings of the European Conference on Computer Vision*, 2016.

[51] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

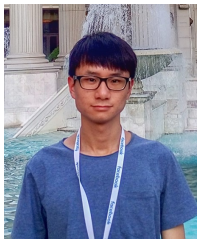


Ji Zhu received the B.E. degree from the School of Electronic Engineering, Xidian University, China, and is currently pursuing the Ph.D. degree at the Department of Electronic Engineering, Shanghai Jiao Tong University, China. He also works as a research scientist at Visbody. His research interests include computer vision, deep learning, and computer graphics.

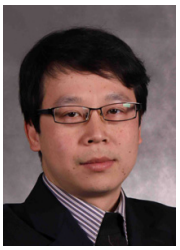
Hua Yang received the Ph.D. degree in communication and information from Shanghai Jiaotong University, in 2004, and both the B.S. and M.S. degrees in communication and information from Haerbin Engineering University, China in 1998 and 2001, respectively. She is currently an associate professor in the Department of Electronic Engineering, Shanghai Jiaotong University, China. She received the first prize of Shanghai technical invention in 2017 and champion of wider person search challenge as an advisor in ECCV2018. Her current research interests include computer vision, machine learning, and smart video surveillance applications.



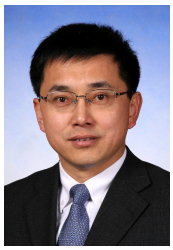
Weiyao Lin received the B.E. and M.E. degrees from Shanghai Jiao Tong University, China, in 2003 and 2005, and the Ph.D degree from the University of Washington, Seattle, USA, in 2010. He is currently a professor at Department of Electronic Engineering, Shanghai Jiao Tong University, China. His research interests include image/video compression, multimedia understanding, and computer vision.



Nian Liu is currently a researcher with the Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE. He received the Ph.D. degree, the M.S. degree, and the B.S. degree from School of Automation at Northwestern Polytechnical University, in 2020, 2015, and 2012, respectively. His research interests include computer vision and machine learning, especially on saliency detection and deep learning.



Jia Wang received the B.Sc. degree in electronic engineering, the M.S. degree in pattern recognition and intelligence control, and the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, China, in 1997, 1999, and 2002, respectively. He is currently a Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University, and also a member of the Shanghai Key Laboratory of Digital Media Processing and Transmission. His research interests include multiuser information theory and mathematics in artificial intelligence.



Wenjun Zhang received his B.S., M.S. and Ph.D. degrees in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 1984, 1987 and 1989, respectively. After three years' working as an engineer at Philips in Nuremberg, Germany, he went back to his Alma Mater in 1993 and became a full professor of Electronic Engineering in 1995. He was one of the main contributors of the Chinese DTTB Standard (DTMB) issued in 2006. He holds 142 patents and published more than 110 papers in international journals and conferences. He is the Chief Scientist of the Chinese Digital TV Engineering Research Centre (NERC-DTV), an industry/government consortium in DTV technology research and standardization, and the director of Cooperative MediaNet Innovation Center (CMIC), an excellence research cluster affirmed by the Chinese Government. His main research interests include video coding and wireless transmission, multimedia semantic analysis and broadcast/broadband network convergence.