Group Reidentification with Multigrained Matching and Integration

Weiyao Lin, Yuxi Li, Hao Xiao, John See, Junni Zou, Hongkai Xiong, Jingdong Wang, and Tao Mei, *Fellow, IEEE*

***Abstract*—The task of reidentifying groups of people under different camera views is an important yet less-studied problem. Group reidentification (Re-ID) is a very challenging task since it is not only adversely affected by common issues in traditional single-object Re-ID problems, such as viewpoint and human pose variations, but also suffers from changes in group layout and group membership. In this paper, we propose a novel concept of group granularity by characterizing a group image by multigrained objects: individual people and subgroups of two and three people within a group. To achieve robust group Re-ID, we first introduce multigrained representations which can be extracted via the development of two separate schemes, that is, one with handcrafted descriptors and another with deep neural networks. The proposed representation seeks to characterize both appearance and spatial relations of multigrained objects, and is further equipped with importance weights which capture variations in intragroup dynamics. Optimal group-wise matching is facilitated by a multiorder matching process which, in turn, dynamically updates the importance weights in iterative fashion. We evaluated three multicamera group datasets containing complex scenarios and large dynamics, with experimental results demonstrating the effectiveness of our approach.**

***Index Terms*—Group reidentification (Re-ID), group-wise matching, multigrained representation, Re-ID.**

I. Introduction

PERSON reidentification (Re-ID) aims at matching and identifying pedestrians across nonoverlapping camera views. This task is increasingly important in visual surveillance and has attracted much attention in recent research [1]–[3]. However, most research focused on individual person Re-ID, while the Re-ID of groups of people is seldom studied. In practice, since most events (e.g., moving, fighting, or violent actions) could be performed within distinct groups instead of between individuals, it is essential to identify groups rather than single people when analyzing events across cameras [4], [5]. Therefore, it is nontrivial to obtain reliable group matching across different camera views.

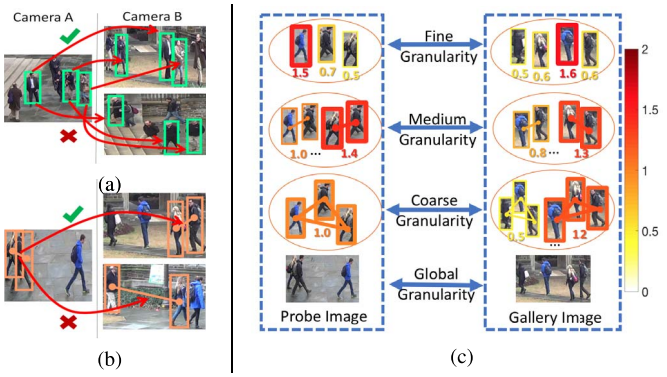


Fig. 1. (a) and (b) Left: Probe groups in camera *A*; right: correctly matched groups (top) and incorrectly matched groups (bottom) in gallery camera *B*. (c) Illustration of multigrained information for group Re-ID. The colored lines and rectangles in (c) indicate the importance weights for individuals and people subgroups. (Best viewed in color.)

In group Re-ID, there are two more basic challenges besides viewpoint changes and human pose variations [3], [6], both inherent issues for the individual person case. These challenges are as follows.

1) *Group Layout Change:* The layout of people in a group is largely unconstrained across different camera views. Due to the dynamic movements of people, the relative positions of people in a group may have large differences in two camera views [see Fig. 1(a)].
2) *Group Membership Change:* People may often join or leave a group [see Fig. 1(b)].

Most existing methods, for example, [4], [5], and [7], view the input group image as an entire unit and extract global/semiglobal features without explicitly performing individual people matching and considering layout changes to perform group-wise matching [4], [5]. A recent study [8] attempts to use descriptors of local patches to partially handle layout and membership changes.

Manuscript received January 26, 2019; revised April 28, 2019 and May 7, 2019; accepted May 12, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61529101, Grant 61425011, and Grant 61720106001, in part by the Shanghai "The Belt and Road" Young Scholar Exchange under Grant 17510740100, and in part by CREST Malaysia under Grant T03C1-17. This paper was recommended by Associate Editor D. Tao. *(Corresponding author: Weiyao Lin.)*

W. Lin, Y. Li, and H. Xiong are with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: wylin@sjtu.edu.cn; lyxok1@sjtu.edu.cn; xionghongkai@sjtu.edu.cn).

H. Xiao is with the Department of Electrical and Computer Engineering, University of Washington, Seattle, WA 98195 USA (e-mail: alexinsjtu@gmail.com).

J. See is with the Faculty of Computing and Informatics, Multimedia University, Cyberjaya 63100, Malaysia (e-mail: johnsee@mmu.edu.my).

J. Zou is with the Department of Computer Science, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: zoujn@cs.sjtu.edu.cn).

J. Wang is with the the Vision Computing Group, Microsoft Research Asia, Beijing 100190, China (e-mail: jingdw@microsoft.com).

T. Mei is with the Laboratory of Computer Vision and Multimedia, JD AI Research, Beijing 101100, China (e-mail: tmei@jd.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2019.2917713

In this paper, we introduce the idea of *group granularity* and characterize a group image by *multigrained objects*. By defining crowds with a large membership overlap from the same group, a group can be depicted with multigrained objects: fine-grained objects are formed by a single person, medium-grained objects are formed by a group of two people, coarse-grained objects are formed by a group of three people, and a global-grained object consists of all people in the group. We argue that characterization of objects from multiple granularities is helpful to enhance the invariance of descriptors toward changes in group membership and layout.

We refer to the example in Fig. 1(a) for better clarity on the need for group granularity. Due to the large layout variation and camera viewpoint changes, the same group shows large global appearance differences in two camera views. The Re-ID performance is poor if global features are merely adopted for the entire group for Re-ID [see the bottom-right image in Fig. 1(a)]. This issue can be resolved if we include information of finer group granularity (e.g., individual persons). On the other hand, merely using the information on individual people is also not always reliable. An example is shown in Fig. 1(b) where two groups in camera *B* contain visually similar group members to the probe group in camera *A*. In this case, the information at medium-level granularity (e.g., subgroups of two people) would be useful.

Our approach leverages on the representations of multiple granularities, also called multigrained objects [see Fig. 1(c)], for group Re-ID. In addition, motivated by the observations that groups in different cameras may be interfered by group member variation, occlusion, and mismatching, and that multigrained objects have different reliabilities on Re-ID performances, we propose introducing the dynamic updated importance weights to explicitly model the different characterization power of each object in different granularities and further improve group Re-ID performance.

Meanwhile, due to the strong ability of convolutional neural networks to extract local invariant features, some deep learning-based methods have achieved unprecedented performance in vision-recognition tasks in recent years [9], [10]. Inspired by these works, many deep learning techniques have been applied to the person Re-ID task [11], [12] and are demonstrated to outperform some traditional pipelines, which are typically reliant on manually designed feature representations and metric-learning algorithms. Nevertheless, few works utilize deep learning methods for group Re-ID. Considering the large variation in illumination, membership change in crowds, and pose transformation in the group Re-ID dataset, there is sufficient motivation to study the performance of deep CNNs on the group Re-ID task.

For comprehensiveness, we introduce two independent pipelines for feature extraction. One is a combination of traditional hand-crafted algorithms while the other is based on a multitask convolutional neural network. Both pipelines extract their own set of appearance and spatial relation features of different granularities. Our experiments convincingly show that our group Re-ID framework is able to achieve state-of-the-art results on different datasets with either handcrafted features or deep convolutional features.

In summary, our contributions are four-fold as follows.

1) We introduce multigrained representations for group images to better handle changes in group layout and membership, coupled with a dynamic weighting scheme for better person matching.
2) We solve the group-wise matching problem by using a multiorder matching algorithm that integrates multigrained representations and combines the information of both matched and unmatched objects to achieve a more reliable matching result.
3) We propose two schemes to extract appearance and spatial relation features for the multigrained representation: one based on typical handcrafted features, and the other based on deep CNN features. For the latter, a new multitask integrated CNN is designed for this specific purpose.
4) We create two challenging group Re-ID datasets with large group membership and layout variations. The existing group Re-ID datasets consist of the relatively small group sizes and group layout, which are less realistic in real-world scenarios.

II. Related Works and Overview

Person Re-ID has been studied for years. Most of them focus on developing reliable features [13], [14]; deriving accurate feature-wise distance metric [15], [16]; and handling local spatial misalignment between people [3], [17]. Some recent research works extend Re-ID algorithms to more object types (e.g., cars [18]) or more complex scenarios (e.g., larger camera numbers [19], long-term videos [20], and untrimmed images [21], [22]).

During recent years, some deep learning-based methods have emerged to solve the problem of single-person Re-ID. These methods have a strong ability to extract rich invariant features from images. A number of works [23], [24] exploit CNNs for person Re-ID by exploiting pairwise labels from positive and negative sample pairs in a variety of network architectures. More recent works [11], [25], [26] are inclined to equip CNNs with triplet loss [27], which has shown to perform exceedingly well [12], [28] by learning a representative feature embedding space which facilitates a distance metric.

However, most existing works focus on the Re-ID of individual person; as such, the group-level Re-ID problem is seldom considered. Since group Re-ID contains significant group layout changes and group membership variations, it introduces new challenges and a proliferation of information that requires encoding compared to scenarios addressed by single-person Re-ID methods. Although some works [29], [30] introduce people or group interaction into the Re-ID process, they only target improving the Re-ID performance of a single person. The characteristics of groups are still less considered and not fully modeled.

Only a few works have been proposed to address group Re-ID tasks [4], [5], [7], [8]. Most of them develop global or semiglobal features to perform group-wise matching. For example, Cai *et al.* [5] proposed a discriminative covariance descriptor to capture the global appearance and statistic

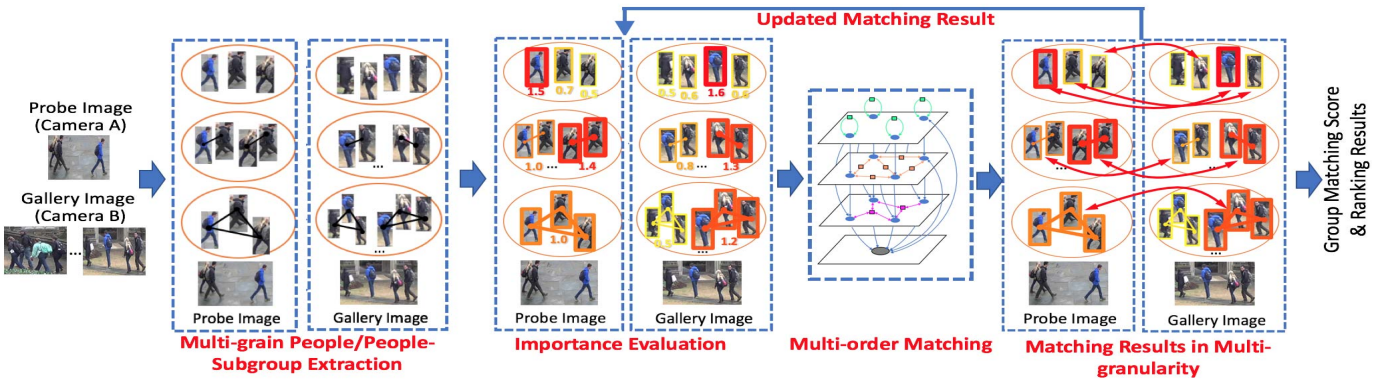


Fig. 2. Framework of the proposed group Re-ID approach, which consists of three sequential parts, that is, multigrained feature extraction, importance evaluation, and multiple order matching. (Best viewed in color.)

properties of group images. Zheng *et al.* [4] segmented a group image into multiple ring regions and derived semiglobal descriptors for each region. Lisanti *et al.* [7] combined sparsity-driven descriptions of all patches into a global group representation. Since global or semiglobal features are unlikely to capture information from local interaction in groups, they may have limitations in handling complex scenarios with significant group appearance variations caused by pose and background interference.

Recently, Zhu *et al.* [8] developed a local-based method which performs group Re-ID by selecting proper patch-pairs and conducting patch matching between cross-view group images. However, in order to reduce patch mismatches, this method includes prior restrictions on vertical misalignments. This limits their capability in handling significant group layout changes or group member variations.

Our approach differs from the existing group Re-ID works in two aspects.

1) The existing works perform Re-ID with information derived from single granularity (i.e., either global or patch-level information). Comparatively, our approach leverages multigrained information to fully capture the characteristics of a group.
2) Our approach does not include any prior restrictions on spatial misalignments, which are able to handle arbitrary group layout changes or group member variations.

Overview of Our Approach: Given the probe group image captured from one camera, our goal is to find the matched group images from a set of gallery group images captured from another camera. We represent each group image by a set of multigrained objects, and then proceed to extract features by a combination of handcrafted descriptors, or by a forward pass on a multitask CNN. With these features, the matching process computes the static and dynamic importance weights of multigrained objects between the probe and gallery images. Then, a multiorder matching algorithm computes intermediate matching results, which are used to update the dynamic importance weights. We perform these two stages in iterative fashion, with final matching results obtained at convergence. The entire framework is shown in Fig. 2.

III. Multigrained Representation

A group image I contains a set of people: $\mathcal{B} = \{b_1, b_2, \ldots, b_N\}$, where N is the number of people and b_i (or simply denoted by i for presentation clarity) corresponds to the person bounding box. The representation is computed by building multigrained objects (people/subgroups): 1) fine granularity, including objects of an individual person, $\mathcal{O}_1 = \{i | i = 1, \ldots, N\}$; 2) medium granularity, including objects of two-people subgroups, $\mathcal{O}_2 = \{(i_1, i_2) | i_1, i_2 = 1, \ldots, N, i_1 \neq i_2\}$; 3) coarse granularity, including objects of three-people subgroups, $\mathcal{O}_3 = \{(i_1, i_2, i_3) | i_1, i_2, i_3 = 1, \ldots, N, i_1 \neq i_2 \neq i_3\}$; and 4) global granularity, referring to the entire group, $\mathcal{O}_g = \{(1, 2, \ldots, N)\}$. In the cases where there are only two people in the group image, we simply let $\mathcal{O}_2$ be the coarse granularity.

Choice of Granularities: We adopt four levels of granularity because the combination of three distinct levels and a global level is sufficient to characterize both the global appearance and local layout of crowds, besides for tractability reasons. The fine granularity helps to reduce the confusion in the global appearance when encountering large layout or group member changes, while the medium and coarse granularities can help to resolve ambiguous individual person matches in the fine granularity by incorporating local layout or co-occurrence information in a group.

Feature Notations: The feature of an object $o \in \mathcal{O}_1$ in the fine granularity, denoted by $\mathbf{f}_o = \mathbf{f}_o^l$, is about the local appearance. The feature of an object $o \in \mathcal{O}_2 \cup \mathcal{O}_3$ in the medium and coarse granularity, denoted by $\mathbf{f}_o = [\mathbf{f}_o^l, \mathbf{f}_o^s]$, consists of two parts: 1) *appearance*, which is an aggregation of features representing the local appearances of all people within the subgroup and 2) *spatial relation*, a single or aggregation of features $\mathbf{f}_o^s$ that describes the spatial layout of each edge linking two different individuals within this subgroup. Here, *aggregation* indicates concatenating the feature of the same granularity and semantics, followed by performing t-SNE [31] for feature reduction. The notation $[\cdot, \cdot]$ denotes a vector concatenation operation, which also applies to the rest of this paper. Meanwhile, the global feature of object $o \in \mathcal{O}_g$, denoted by $\mathbf{f}_o = \mathbf{f}_o^g$, describes the appearance feature of the entire input group image.

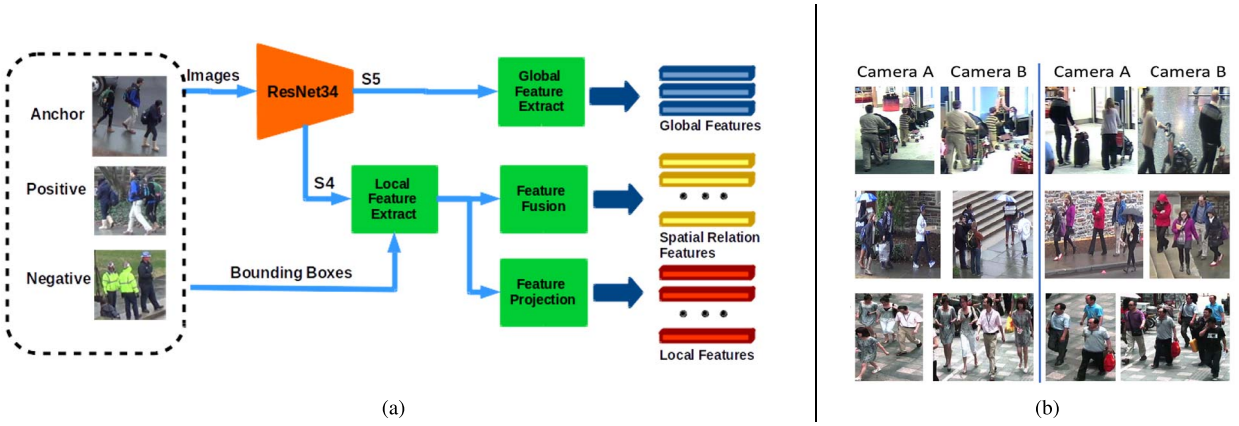


Fig. 3. (a) Overview of our multitask CNN for feature extraction from input group images. (Best viewed in color.) The dotted rectangle denotes a triplet of input images consisting of the anchor, positive, and negative samples. S4 and S5 denote the fourth and fifth stage of ResNet-34 CNN. (b) Examples from datasets used in our experiments. The first row is from *i-LID MCTS*. The second and third rows are from our constructed *DukeMTMC Group* and *Road Group* datasets, respectively.

As shown in Fig. 2, our framework is independent of the choice of feature vectors used, which implies that we could exploit some conventional algorithms to extract handcrafted features. On the other hand, we also intend to exploit the strong ability of deep CNNs in extracting more representative and invariant features. Therefore, we introduce two different pipelines to obtain these feature representations for usage in our group Re-ID framework. The first extracts a combination of manually designed descriptors to encode both object appearances and layout relationship between objects while the second is a new integrated deep learning-based method to extract features in a single forward pass.

A. Handcrafted Feature Descriptors

In this section, we briefly describe the handcrafted features utilized in this paper. Color and texture features [15] are used as the *appearance* part of the object. To be specific, for each single person input image (obtained from bounding box b_i and resized to unified resolution), we split it into 18 equal-sized patches along the vertical direction, and the RGB, HSV, YCbCr, LAB, YIQ color features, and Gabor texture features are extracted from each patch. The output-normalized histograms of these features are concatenated to form a final 8024-D local *appearance* descriptor $\mathbf{f}_o^l$. Similarly, we also extract global features $\mathbf{f}_o^g$ in the same manner but with the input image containing the entire group.

As for the *spatial relation* part, we use the relative distance and angle histograms among individuals in an object [32] to describe each edge between two people (i, j). For two bounding boxes b_i and b_j within a subgroup, we first denote the relative position between their centers in polar coordinate (ρ_{ij}, θ_{ij}), where ρ_{ij} is the log-distance between the two centers and θ_{ij} is the corresponding orientation angle. The 10-D log-distance histogram L_{ij} and 9-D angle histogram P_{ij} are constructed over uniform bins as follows:

$$L_{ij}(k) = \mathcal{N}(k - m; 0, \sigma_L) \quad (1)$$

$$P_{ij}(k) = \mathcal{N}\big(k - m_{ij}; 0, \sigma_P\big) + \mathcal{N}\big(k - m_{ij}; \pm 9, \sigma_P\big) \quad (2)$$

where $\mathcal{N}(x; \mu, \sigma)$ is a discrete Gaussian window parameterized by mean μ and variance σ, while m_{ij} is the index of the bin containing ρ_{ij} or θ_{ij}. Finally, the two output histograms are combined to form the descriptor that represents the b_i–b_j edge

$$\mathbf{f}^s_{(i,j)} = [L_{ij}, P_{ij}]. \quad (3)$$

B. Integrated CNN for Feature Extraction

Inspired by modern CNN-based object detection frameworks [33], which perform exceedingly well in both recognition and localization tasks, we hypothesize that deep neural networks can be tailored to handle both the appearance and structural layout of objects in an integrated manner. As such, we propose a new multitask network to jointly extract both *appearance* and *spatial relation* features needed for our group Re-ID framework.

The overview of the multiple-task CNN in this paper is depicted in Fig. 3(a). For each input image, we use the ResNet-34 [9] as the backbone structure for basic feature extraction. Postprocessing includes two separate branches: 1) the global branch, which is responsible for extracting features from the entire group image and 2) a local branch, which is utilized to handle individual objects and their relations to others. With these branches processed in parallel, we could obtain the multigrained features simultaneously.

1) Minibatch Organization: We borrow the idea of using triplet loss for training [11], [12], [34] to learn representative mapping of images to an abstract feature space. Therefore, we organize the minibatch into triplets where each training sample consists of three images: 1) an anchor image I_a from probe images; 2) a positive image I_p from the gallery which contains the same group as the anchor; and 3) a negative image I_n which is randomly selected from the training set that has a different group id from the anchor and positive images.

Throughout this section, we shall adopt the same subscripts to denote anchor, positive, and negative images for features or object sets. For example, $\mathbf{f}_a^g$ denotes the global feature of the anchor image and $\mathcal{O}_{1,a}$ denotes the set of person objects in the anchor image.

2) Global Feature Extraction: The global branch receives feature maps from the final convolution layer of the fifth stage of ResNet-34, which is denoted as S5 in Fig. 3(a). We then apply a simple global average pooling operation followed by a fully connected layer as the global feature extraction module to obtain the corresponding global features $\mathbf{f}_m^g$, $m \in \{a, p, n\}$.

3) Local Feature Extraction: The local branch receives feature maps from the final convolutional layer of the fourth stage of ResNet-34 [denoted as S4 in Fig. 3(a)] and bounding box sets $\mathcal{B}_a$, $\mathcal{B}_p$, and $\mathcal{B}_n$, belonging to each image of the triplet. The local feature extraction module applies ROI Pooling [33] on each feature map according to its bounding box and sends the output to a fully connected layer. Hence, we obtain the intermediate local features $\hat{\mathbf{f}}_{i,m}^l$ for each ith individual person in the image of set $m \in \{a, p, n\}$.

Next, these features are further utilized in two sub-branches. First, a feature projection module takes the intermediate features as input and yields local appearance features by a nonlinear projection that constrains outputs to $(-1, 1)$

$$\mathbf{f}_{i,m}^l = \tanh\left(\mathbf{W}_l \hat{\mathbf{f}}_{i,m}^l\right) \tag{4}$$

where $\mathbf{W}_l$ is a learnable transformation matrix. Second, to model the spatial relation between two objects, that is, an edge linking two individuals (i, j), a feature fusion module is applied to fuse intermediate features with another learnable transformation matrix $\mathbf{W}_s$ as follows:

$$\mathbf{f}_{(i,j),m}^s = \tanh\left(\mathbf{W}_s\left[\hat{\mathbf{f}}_{i,m}^l, \hat{\mathbf{f}}_{j,m}^l\right]\right). \tag{5}$$

4) Loss Function: To regularize the global and local output features from different branches, we design three different types of losses to train our multitask CNN. For the global appearance features, we intend to learn an embedding space where the anchor sample will be closer to the positive sample than the negative sample. For this, we obtain a group-wise training loss by utilizing the triplet loss

$$L_g = \left[d_2\left(\mathbf{f}_a^g, \mathbf{f}_p^g\right) - d_2\left(\mathbf{f}_a^g, \mathbf{f}_n^g\right) + \lambda_g\right]_+ \tag{6}$$

where $[\cdot]_+$ is the ReLU operator $\max(0, \cdot)$, $d_2(\cdot, \cdot)$ denotes the L2-norm distance between the two feature vectors, and λ_g is a hyperparameter to control the margin size.

For the supervision of individual appearance features, we could simply apply a triplet loss over the matched and unmatched pairs, similar to that in (6). However, since the number of matched individual pairs between anchor and positive images is much less than that of unmatched ones, the loss might well be dominated by unmatched pairs. Inspired by the Trihard loss [12], we impose the hard negative mining strategy to the standard triplet loss, much akin to a k-nearest negative neighbors manner

$$L_a = \frac{1}{|\mathcal{O}_{1,a}|} \sum_{i \in \mathcal{O}_{1,a}} \left[d_2\left(\mathbf{f}_{i,a}^l, \mathbf{f}_{i',p}^l\right) - d_{\text{neg}}(i) + \lambda_l\right]_+ \tag{7}$$

where $i' \in \mathcal{O}_{1,p}$ is the individual person in the positive image who matches exactly to the ith person in the anchor image. Note that if person i in anchor does not match any individual in the positive image, we set the $d_2(\cdot, \cdot)$ term in (7) to be zero. The hyperparameter λ_l controls the margin size of local appearance features. The term $d_{\text{neg}}(i)$ is the average distance between person i in the anchor image and people in set $\mathcal{K} = \mathcal{K}_p \cup \mathcal{K}_n$, which carries the intuition of the k-nearest unmatched individuals from other images in the feature space

$$d_{\text{neg}}(i) = \frac{1}{|\mathcal{K}|}\left(\sum_{j \in \mathcal{K}_p} d_2\left(\mathbf{f}_{i,a}^l, \mathbf{f}_{j,p}^l\right) + \sum_{j \in \mathcal{K}_n} d_2\left(\mathbf{f}_{i,a}^l, \mathbf{f}_{j,p}^l\right)\right) \tag{8}$$

where $\mathcal{K}_p$ and $\mathcal{K}_n$ denote the collection of k-nearest unmatched individuals from the positive and negative images, respectively.

In the work of [33], deep neural networks can precisely predict the relative offset between two bounding boxes. Based on this observation, we design a regression loss to supervise the learning of spatial relation features. Given the bounding boxes, $b_{i,m}, b_{j,m} \in \mathcal{B}_m$ (with $m \in \{a, p, n\}$ as mentioned before) of two individuals (i, j) in an image and the feature $\mathbf{f}_{(i,j),m}^s$ representing the edge between them, we apply linear regression with learnable parameter W_r to predict the normalized spatial transition

$$\left[\hat{\delta}_{(i,j),m}^x, \hat{\delta}_{(i,j),m}^y\right]^T = W_r \mathbf{f}_{(i,j),m}^s. \tag{9}$$

Suppose the bounding box is in the form of $b_{m,i} = (x_{i,m}, y_{i,m}, w_{i,m}, h_{i,m})$ as defined in [33], then we denote the ground-truth transition as

$$\delta_{(i,j),m}^x = \frac{x_{j,m} - x_{i,m}}{w_{i,m}} \tag{10}$$

$$\delta_{(i,j),m}^y = \frac{y_{j,m} - y_{i,m}}{h_{i,m}}. \tag{11}$$

Hence, the localization loss over the spatial relation features is

$$L_s = \frac{1}{P} \sum_{m \in \{a,p,n\}} \sum_{(i,j) \in \mathcal{O}_{2,m}} \sum_{t \in \{x,y\}} S\left(\hat{\delta}_{(i,j),m}^t, \delta_{(i,j),m}^t\right) \tag{12}$$

where $P = \sum_{m \in \{a,p,n\}} |\mathcal{O}_{2,m}|$ is the sum of the total number of bounding box pairs within each image of the triplet. $S(\cdot, \cdot)$ denotes the smooth L1 loss used also in [33]. Finally, we combined the losses in (6), (7), and (12) to obtain the final objective function for our multitask CNN framework

$$L = L_g + \lambda_1 L_a + \lambda_2 L_s \tag{13}$$

where λ_1 and λ_2 are balancing factors. After the training phase, this multitask network is used to extract deep multigrained features in a single forward pass. Aggregation of these local features is performed to attain features of coarse and medium granularities for the subsequent matching step. In our experiments, we set λ_g and λ_l as 2.0, and the balancing factors λ_1 and λ_2 as 1.0 during training.

IV. Importance Weighting

We introduce an importance weight α_o for each object o (except the global-grained object) to indicate the object's discriminativity and reliability inside the group image for group person matching. The importance weighting scheme is partially inspired by but different from the saliency-learning methods [6], [8] for differentiating *patch* reliabilities in person Re-ID: 1) our scheme aims to weigh each granularity object

rather than patches and 2) our scheme dynamically adjusts the importance weights in an iterative manner, by using the intermediate matching results at each iteration (see Fig. 2).

A. Fine-Grained Object

The importance weight (α_i) for each individual person i in the probe image I consists of two components: 1) static weight, which is only dependent on the group image and 2) dynamic weight, which is dynamically updated according to the intermediate matching results with the gallery group images, from another camera in our approach. The formulation of this weight is given as follows:

$$\alpha_i = t_1\left(i, \mathcal{G}_{\setminus i}\right) + s(i, \mathcal{M}_i) + p\left(\mathcal{M}_i, \mathcal{M}_{\mathcal{G}_{\setminus i}}\right) \tag{14}$$

where the first term is the static weight, and the second and third terms form the dynamic weight.

Static Weight: The static weight $t_1(i, \mathcal{G}_{\setminus i})$, where $\mathcal{G}_{\setminus i} = \mathcal{G} - \{i\}$ denotes the set of other individual people in $\mathcal{G}$, is used to describe the stability. It is computed as follows:

$$t_1\left(i, \mathcal{G}_{\setminus i}\right) = \sum_{i' \in \mathcal{G}_{\setminus i}} \frac{\rho_i}{\rho_{i'}} \tag{15}$$

where ρ_i is the local density around person i in group $\mathcal{G}$. It reflects the density of people in a neighborhood around i, which is computed by following [35].

By (15), the static weight t_1 is mainly obtained by evaluating the relative local density ratios between person i and his/her peer group members i' in $\mathcal{G}$. If the local density around i is larger than the density around his/her peer group members i', the stability of i is increased, indicating that i is located in the *center* region of group $\mathcal{G}$ and should be a more reliable member in group Re-ID [see person 1 in Fig. 4(a)]. On the contrary, when i's local density is smaller than his/her peer group members, a small stability value will be assigned, indicating that i is located in the outlier region of the group and is less reliable [see person 2 in Fig. 4(a)].

Dynamic Weight: The dynamic weight $s(i, \mathcal{M}_i) + p(\mathcal{M}_i, \mathcal{M}_{\mathcal{G}_{\setminus i}})$ consists of two parts: 1) the saliency term $s(i, \mathcal{M}_i)$ and 2) the purity term $p(\mathcal{M}_i, \mathcal{M}_{\mathcal{G}_{\setminus i}})$, where $\mathcal{M}_i$ is the set of matches from the gallery group images, and $\mathcal{M}_{\mathcal{G}_{\setminus i}}$ is the set of matches for all people except i in the probe image, $\mathcal{M}_{\mathcal{G}_{\setminus i}} = \{\mathcal{M}_{i'} | i' \notin \mathcal{G}\}$. The sets of matches are illustrated in Fig. 4(a).

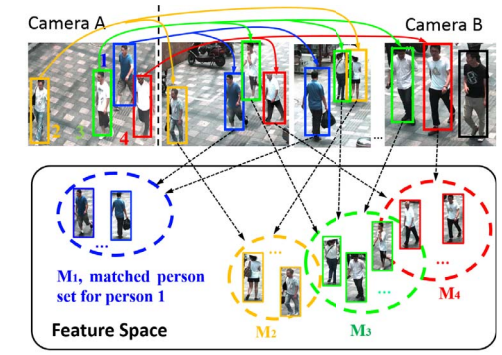
The saliency term is computed as

$$s(i, \mathcal{M}_i) = \lambda_s \frac{d_f\left(\mathbf{f}_i, \mathbf{f}_{\mathcal{M}_i}\right)}{|\mathcal{M}_i|}. \tag{16}$$

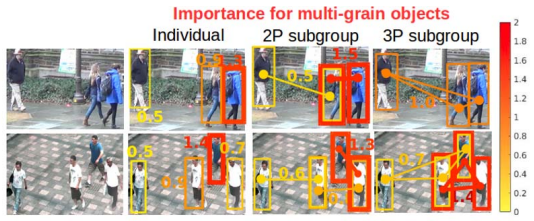
Here, $d_f(\cdot)$ is the Euclidean distance between features. $|\mathcal{M}_i|$ is the cardinality of $\mathcal{M}_i$. $\mathbf{f}_{\mathcal{M}_i}$ is the feature describing the set of matches $\mathcal{M}_i$, and we use the feature of the $(1/2)|\mathcal{M}_i|$th nearest neighbor of i in $\mathcal{M}_i$ as done in [6] and [36]. λ_s is an adaptive normalization factor to normalize the range of s to be within 0 to 1, which is computed as follows:

$$\lambda_s = \frac{1}{\sum_i d_f\left(\mathbf{f}_i, \mathbf{f}_{\mathcal{M}_i}\right)/|\mathcal{M}_i|}. \tag{17}$$

For simplicity, the other normalization factors in the rest of this paper are calculated in the similar way as λ_s.



(a)



(b)

Fig. 4. (a) Illustration of matched-people sets and their distributions in the feature space. (The color solid arrows indicate the one-to-one mapping results between individuals. People circled by the same color rectangles in camera *B* are matched to the same person in *A*, and belong to the same matched-people set.) (b) Derived importance weights for multigrained objects (individuals, 2-people subgroups, and 3-people subgroups) in two example group images. Note: the importance weights for some 2-people/3-people subgroups are not displayed in order for a clear illustration. (Best viewed in color.)

According to (16), if the appearance of an individual person i is discriminative, a large portion of individuals in i's matched set $\mathcal{M}_i$ are visually dissimilar to i. This leads to a large $d_f(\mathbf{f}_i, \mathbf{f}_{\mathcal{M}_i})$ and a large saliency value [6], [36] [see person 1 in Fig. 4(a)]. Moreover, due to the variation of group members in group Re-ID, each individual person may have a different number of matched people in his/her $\mathcal{M}_i$. Therefore, we further introduce $|\mathcal{M}_i|$ in (16) such that the person with fewer matched people can indicate a more discriminative appearance.

The purity term is computed as

$$p\left(\mathcal{M}_i, \mathcal{M}_{\mathcal{G}_{\setminus i}}\right) = \sum_{i' \in \mathcal{G}_{\setminus i}} \lambda_p d_m(\mathcal{M}_i, \mathcal{M}_{i'}) \tag{18}$$

where $d_m(\cdot)$ is the Wasserstein-1 distance [37], a measure to evaluate the dissimilarity between two feature sets. λ_p is calculated in the same way as λ_s in (16).

According to (18) and Fig. 4(a), the purity measurement reflects the relative appearance uniqueness of person i inside group $\mathcal{G}$. If i has similar appearance features as other group members in $\mathcal{G}$, their matched people in camera B should also be visually similar and located close to each other in the feature space [see $\mathcal{M}_3$ and $\mathcal{M}_4$ in Fig. 4(a)], resulting in a small purity value. On the other hand, if a person includes *unique* appearance features in $\mathcal{G}$, his/her matched people in camera B should have larger feature distances than those of the other members in $\mathcal{G}$, and lead to a large purity value [see $\mathcal{M}_1$ in Fig. 4(a)].

B. Medium and Coarse-Grained Objects

The importance weight $\alpha_{i_1 i_2}$ of a medium-grained object (i_1, i_2) is computed as

$$\alpha_{i_1 i_2} = \alpha_{i_1} + \alpha_{i_2} + t_2(i_1, i_2). \tag{19}$$

Here, $t_2(i_1, i_2)$ is the stability measure of the subgroup (i_1, i_2). A two-person subgroup is thought to be more stable if its members are spatially closer to each other. Thus, we simply compute t_2 by the inverse of spatial distance between i_1 and i_2.

The importance weight $\alpha_{i_1 i_2 i_3}$ of a coarse-grained object (i_1, i_2, i_3) is computed as

$$\alpha_{i_1 i_2 i_3} = \alpha_{i_1 i_2} + \alpha_{i_2 i_3} + \alpha_{i_1 i_3} + t_3(i_1, i_2, i_3). \tag{20}$$

Here, $\alpha_{i_1 i_2}$ is the importance of a two-people pair in (i_1, i_2, i_3) [see (19)]. t_3 is the stability of a three-person subgroup. We assume the equilateral triangle as the most stable structure for three-people subgroups and model t_3 by evaluating its similarity to an equilateral triangle according to (21), where $\theta_k, k \in \{1, 2, 3\}$ denote the angles of the triangle constructed by coarse subgroup (i_1, i_2, i_3)

$$t_3(i_1, i_2, i_3) = \exp\left(-2 * \sum_{k=1}^{3} |\sin\theta_k - \sin\frac{\pi}{3}|\right). \tag{21}$$

Fig. 4(b) shows the importance weights of some groups. From Fig. 4(b), we can see that our process can effectively set larger weights on objects with stronger characterization ability to represent the entire group.

C. Iterative Update

We utilize an iterative process which updates the importance weights and group-wise matching results iteratively. We initialize the dynamic weights for all objects by 1 and compute the optimal matching through multiorder matching (see Section V) to obtain an initial matching result: $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_N$. This matching result is used to update the dynamic importance weights. This procedure is repeated until the importance weights become converged or the maximum iteration is reached. Although the exact convergence of our iterative process is difficult to analyze due to the inclusion of multiorder matching, in our experiments, we confirm that most important weights become stable within five iterations, which implies the reliability of our approach.

V. Multiorder Matching

Given a probe image I_p and a gallery image I_g, our goal is to compute the matching score between the two groups of people. Suppose that there are N_p people in the probe image, and I_p and N_g people in the gallery image I_g. The goal of the multiorder matching process aims to find: 1) an optimal one-to-one mapping, $\mathcal{C} = \{(i, j)|\ \forall (i, j), (i', j'), i \neq i', j \neq j'\}$, where (i, j) $(= c_{ij})$ denotes a match between the ith person from the probe image and the jth person from the gallery image and 2) the maximum matching score.

Since a group is characterized by multiple granularities, it is natural to measure the similarity across different granularities to find the optimal match. With this consideration, the objective function of our matching process is formulated with multiorder potentials

$$\mathcal{Q}(\mathcal{C}) = \mathcal{P}_1(\mathcal{C}) + \mathcal{P}_2(\mathcal{C}) + \mathcal{P}_3(\mathcal{C}) + \mathcal{P}_g(\mathcal{C}) + \sum_{r \neq l, r, l \in \{1,2,3,g\}} \mathcal{P}_{rl}(\mathcal{C}) \tag{22}$$

where $\mathcal{P}_1(\mathcal{C})$, $\mathcal{P}_2(\mathcal{C})$, $\mathcal{P}_3(\mathcal{C})$, and $\mathcal{P}_g(\mathcal{C})$ are the first-order, second-order, third-order, and global potentials, evaluating the matching quality over each subgroup of people, and $\mathcal{P}_{rl}(\mathcal{C})$ is the interorder potential.

A. Multiorder Potentials

First-Order Potential: $\mathcal{P}_1(\mathcal{C})$ is used to model the matching scores over individual people. It is calculated by the sum of the matching scores of all the individual matches in $\mathcal{C}$

$$\mathcal{P}_1(\mathcal{C}) = \sum_{c_{ij} \in \mathcal{C}} m_1(c_{ij}) = \sum_{c_{ij} \in \mathcal{C}} w_1(\mathbf{f}_i, \alpha_i, \mathbf{f}_j, \alpha_j) \tag{23}$$

where $\mathbf{f}_i$, α_i and $\mathbf{f}_j$, α_j are the feature vector and importance weight for the probe-image person i and gallery-image person j, respectively [see (14)]. $m_1(c_{ij}) = w_1(\mathbf{f}_i, \alpha_i, \mathbf{f}_j, \alpha_j)$ is the matching score for match $c_{ij} = (i, j)$, calculated by

$$m_1(c_{ij}) = w_1(\mathbf{f}_i, \alpha_i, \mathbf{f}_j, \alpha_j) = \lambda_{w_1} \frac{\psi(\alpha_i, \alpha_j)}{d_{\mathbf{f}}(\mathbf{f}_i, \mathbf{f}_j)} \tag{24}$$

where $\psi(\alpha_i, \alpha_j) = [(\alpha_i + \alpha_j)/(1 + |\alpha_i - \alpha_j|)]$ is the fused importance weight, which will have a large value if the importance weights of α_i and α_j are both large and close to each other. $d_{\mathbf{f}}(\cdot)$ is the Euclidean distance and λ_{w_1} is the normalization constant for the first-order potential.

By (24), the matching score $m_1(c_{ij})$ is computed by the importance-weighted feature similarity $w_1(\mathbf{f}_i, \alpha_i, \mathbf{f}_j, \alpha_j)$ between the matched individuals i and j.

Second-Order Potential: $\mathcal{P}_2(\mathcal{C})$ is used to model the matching scores over two-people subgroups

$$\begin{aligned}\mathcal{P}_2(\mathcal{C}) &= \sum_{c_{i_1 j_1}, c_{i_2 j_2} \in \mathcal{C}} m_2(c_{i_1 j_1}, c_{i_2 j_2}) \\ &= \sum_{c_{i_1 j_1}, c_{i_2 j_2} \in \mathcal{C}} w_2(\mathbf{f}_{i_1 i_2}, \alpha_{i_1 i_2}, \mathbf{f}_{j_1 j_2}, \alpha_{j_1 j_2})\end{aligned} \tag{25}$$

where $\mathbf{f}_{i_1 i_2}$, $\alpha_{i_1 i_2}$ and $\mathbf{f}_{j_1 j_2}$, $\alpha_{j_1 j_2}$ are the feature vector and importance weight for the probe-image subgroup (i_1, i_2) and gallery-image subgroup (j_1, j_2), respectively [see (19)]. $m_2(c_{i_1 j_1}, c_{i_2 j_2}) = w_2(\mathbf{f}_{i_1 i_2}, \alpha_{i_1 i_2}, \mathbf{f}_{j_1 j_2}, \alpha_{j_1 j_2})$ is the second-order match score between two-people subgroups (i_1, i_2) and (j_1, j_2), which is calculated in a similar way as (24)

$$w_2(\mathbf{f}_{i_1 i_2}, \alpha_{i_1 i_2}, \mathbf{f}_{j_1 j_2}, \alpha_{j_1 j_2}) = \lambda_{w_2} \frac{\psi(\alpha_{i_1 i_2}, \alpha_{j_1 j_2})}{d_{\mathbf{f}}(\mathbf{f}_{i_1 i_2}, \mathbf{f}_{j_1 j_2})}. \tag{26}$$

Third-Order Potential: $\mathcal{P}_3(\mathcal{C})$ is used to model the matching scores over three-people subgroups

$$\begin{aligned}\mathcal{P}_3(\mathcal{C}) &= \sum_{c_{i_1 j_1}, c_{i_2 j_2}, c_{i_3 j_3} \in \mathcal{C}} m_3(c_{i_1 j_1}, c_{i_2 j_2}, c_{i_3 j_3}) \\ &= \sum_{c_{i_1 j_1}, c_{i_2 j_2}, c_{i_3 j_3} \in \mathcal{C}} w_3(\mathbf{f}_{i_1 i_2 i_3}, \alpha_{i_1 i_2 i_3}, \mathbf{f}_{j_1 j_2 j_3}, \alpha_{j_1 j_2 j_3})\end{aligned} \tag{27}$$

where the term $w_3(\mathbf{f}_{i_1i_2i_3}, \alpha_{i_1i_2i_3}, \mathbf{f}_{j_1j_2j_3}, \alpha_{j_1j_2j_3}) = m_3(c_{i_1j_1}, c_{i_2j_2}, c_{i_3j_3})$ is the third-order match score between three-people subgroups (i_1, i_2, i_3) and (j_1, j_2, j_3). It is calculated in the same way as (26).

Global Potential: The global potential is calculated by the global matching score between probe and gallery images I_p and I_g

$$\begin{aligned}\mathcal{P}_g(\mathcal{C}) &= \sum_{\mathcal{C}} m_g\left(c_{i_1j_1}, c_{i_2j_2}, \ldots, c_{i_{N_p}j_{N_q}}\right) \\ &= w_g\left(\mathbf{f}_p, \alpha_p, \mathbf{f}_g, \alpha_g\right)\end{aligned} \tag{28}$$

where $\mathbf{f}_p$ and $\mathbf{f}_g$ are the global feature vectors for the entire group images I_p and I_g. $\alpha_p = \alpha_g = 1$ are the importance weights for global objects. In this paper, we simply use the global feature similarity as the global matching score, as $w_g(\mathbf{f}_p, \alpha_p, \mathbf{f}_g, \alpha_g) = [1/(d_{\mathbf{f}}(\mathbf{f}_p, \mathbf{f}_g))]$.

Interorder Potential: Since each match c_{ij} is described by potentials in multiple orders [see (23)–(28)], we also introduce interorder potentials to properly combine this multiorder potential information. Specifically, the interorder potential between orders $r, l \in \{1, 2, 3, g\}$ is calculated by

$$\mathcal{P}_{rl}(\mathcal{C}) = \sum_{c_{ij} \in \mathcal{C}} m_{rl}\left(c_{ij}, r, l\right) \tag{29}$$

where $m_{rl}(c_{ij}, r, l)$ is the interorder correlation for match c_{ij}. It is calculated by

$$\begin{aligned} m_{rl}\left(c_{ij}, r, l\right) &= \frac{\overline{m}_r\left(c_{ij}, r\right) + \overline{m}_l\left(c_{ij}, l\right)}{1 + \left|\overline{m}_r\left(c_{ij}, r\right) - \overline{m}_l\left(c_{ij}, l\right)\right|} \\ \text{for } \overline{m}_k\left(c_{ij}, k\right) &= \lambda_k \sum_{c_{i'_1j'_1} = c_{ij}} m_k\left(c_{i'_1j'_1} \ldots c_{i'_kj'_k}\right) \end{aligned} \tag{30}$$

where λ_k is the normalization constant for order k, and $m_k(c_{i'_1j'_1} \ldots c_{i'_kj'_k})$ is the intraorder match score in order k [as in (24) and (26)]. From (30), if a match c_{ij} creates large and similar intraorder match scores in both the rth and lth order, it will be considered as being more valuable and reliable and, thus, will have larger interlevel potentials.

B. Optimization

The objective function in (22) properly integrates the information of multigrained objects. Thus, by maximizing (22), we are able to obtain the optimal one-to-one mapping result among individuals in probe and gallery groups.

To solve the multiorder matching problem in (22), we construct a multiorder association graph to incorporate all candidate matches and multiorder potentials in the objective function, as in Fig. 5. In Fig. 5, each layer includes all candidate matches c_{ij} (the circular nodes) and their corresponding intraorder matching scores m_k (the rectangular nodes in green, orange, or pink), which models the intraorder potentials in a specific order. Besides, the blue rectangular nodes linking circular nodes in different layers represent the interorder correlations $m_{rl}(c_{ij}, r, l)$. They model the interorder potentials between different orders.

With this association graph, we are able to solve (22) by adopting general-purpose hypergraph-matching solvers [38], [39]. Specifically, we first initialize a mapping probability for each candidate match in the association graph, and then apply a reweighted random walk [39] to update these mapping probabilities via the interorder/intraorder links and potential weights in the association graph. Finally, the mapping probabilities in all layers in the association graph are combined to obtain the optimal one-to-one mapping result from the candidate matches [38].

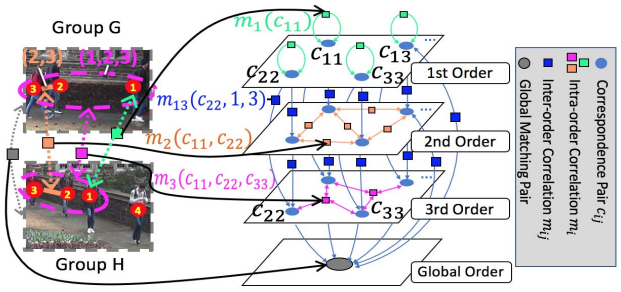


Fig. 5. Illustration of the multiorder association graph. Left: Cross-view group pair being matched. Right: Multiorder association graph constructed for the group pair. (Best viewed in color.)

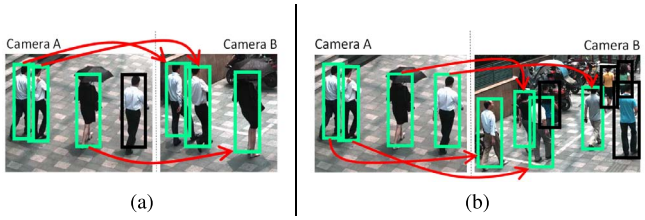


Fig. 6. Illustration of the unmatched term in (31). (a) True match pair. (b) False match pair. Green and black rectangles show matched and unmatched individuals, respectively. Since the right group in (b) includes more individuals, we can find more matched pairs. This may misleadingly result in a high similarity score. However, when considering the large number of unmatched people in (b), the matching score of (b) ought to be properly reduced.

C. Fused Matching Score

After obtaining one-to-one mapping between individual people in two groups, we are able to calculate matching scores accordingly. In order to obtain a more reliable matching score, we introduce a fused scheme by integrating the information of both matched and unmatched objects

$$\begin{aligned} S(I_p, I_g) = &\sum_k \sum_{(i_1 \ldots i_k) \in \mathcal{R}_p} \frac{w_k\left(\mathbf{f}_{i_1 \ldots i_k}, \alpha_{i_1 \ldots i_k}, \mathbf{f}_{M(i_1 \ldots i_k)}, \alpha_{M(i_1 \ldots i_k)}\right)}{|\mathcal{R}_p|} \\ &- \lambda_r \cdot \sum_k \left(\sum_{(i_1 \ldots i_k) \in \overline{\mathcal{R}}_p} \frac{a_{i_1 \ldots i_k}}{|\overline{\mathcal{R}}_p|} + \sum_{(j_1 \ldots j_k) \in \overline{\mathcal{R}}_g} \frac{a_{j_1 \ldots j_k}}{|\overline{\mathcal{R}}_g|} \right) \end{aligned} \tag{31}$$

where $(i_1, \ldots, i_k)$ is a person/subgroup in a probe group image I_p, and $M(i_1 \ldots i_k)$ is its one-to-one matched person/subgroup in gallery image I_g. $w_k(\cdot)$ is the similarity matching score between $(i_1, \ldots, i_k)$ and $M(i_1 \ldots i_k)$, as in (24) and (26). α is the importance weight. $\lambda_r = 0.5$ is a balancing factor. $\mathcal{R}_p$ and $\mathcal{R}_g$ are the sets of reliably matched objects in groups I_p and I_g, and $\overline{\mathcal{R}}_p$ and $\overline{\mathcal{R}}_g$ are the unmatched object sets. The

TABLE I
STATISTICAL ANALYSIS OF DATASETS

Datasets	*i-LID MCTS*	*Road Group*	*DukeMTMC Group*
Average Person	2.313	3.812	3.392
Member Change	0.187	0.451	0.832
Dispersity	0.317	0.376	0.407
Occlusion Pairs	1.021	1.775	2.001

matched object pairs that maximize the objective function (22) are taken as the reliably matched objects, and put into $\mathcal{R}_p$ and $\mathcal{R}_g$. The remaining unmatched or fewer similar objects are put into $\overline{\mathcal{R}}_p$ and $\overline{\mathcal{R}}_g$.

From (31), our fused scheme integrates four granularities (i.e., $k = 1, 2, 3, g$) to compute the group-wise matching score. Inside each granularity, the matched pairs that maximize (22) are used to compute the similarity [the first term in (31)] in order to reduce the interference of confusing or mismatched people/people subgroups. Meanwhile, we introduce an unmatched term evaluating the importance of unmatched objects [the second term in (31)]. As such, we can properly avoid misleadingly high matching scores in false group pairs (as in Fig. 6) and obtain a more reliable result.

VI. EXPERIMENTAL RESULTS

We perform experiments on three datasets: 1) the publicly available *i-LID MCTS* dataset [4] which contains 274 group images for 64 groups; 2) our newly constructed *DukeMTMC Group* dataset which includes 177 group image pairs extracted from the 8-camera-view DukeMTMC dataset [40]; and 3) our newly collected *Road Group* dataset which consists of 162 group pairs taken from a 2-camera crowded road scene.[1]

To construct the *Road Group* dataset, we use [41] to automatically identify groups from key frames that were extracted at equi-intervals of 50 frames. Then, the group image pairs are randomly selected from sets according to different group sizes and occlusion variations. We define two cross-view images from different cameras as belonging to the same group when they have more than 60% of members in common.

Some example of groups from the three datasets are shown in Fig. 3(b), showing diverse challenging conditions across cameras. We also provide a statistical analysis of the three datasets in Table I. "Average Person" denotes the average number of individuals per group image, while "Member Change" denotes the average difference in group size for each pair of group images. "Dispersity" is measured by averaging the normalized distance between each individual member to its group centroid. This measure computes the sparsity (or compactness) of the group, which indirectly indicates the proneness to layout change. "Occlusion Pairs" denotes the average number of individual pairs that occlude each other within an image. Note that despite the *i-LID MCTS* dataset having smaller and more compact groups, it suffers from low-image quality and large illumination changes. Meanwhile, the new datasets *DukeMTMC Group* and *Road Group* are both plagued with severe object occlusions, and large layout and group member variations due to larger group sizes.

[1]Dataset and source code will be available at http://min.sjtu.edu.cn/lwydemo/GroupReID.html.

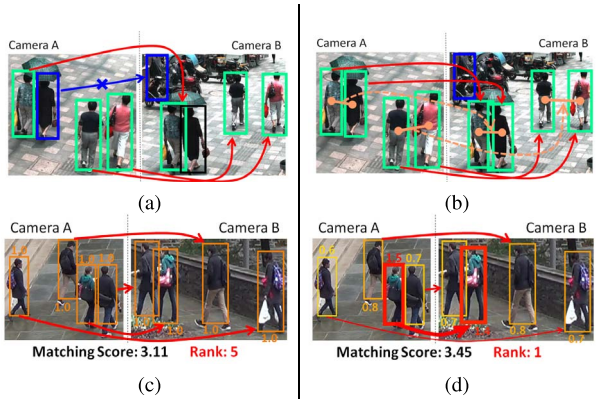


Fig. 7. Matching results by: (a) using only information of individuals; (b) using multigrained information; (c) setting equal importance weights for all individuals/subgroups; and (d) using our importance evaluation process to obtain importance weights. The red and blue links indicate correct and wrong matches, respectively. (Note: For a clear illustration, we only display the matching results between individuals.)

A. Experimental Setup

To provide a fair comparison with other methods, we follow the evaluation protocol in [8] and [42] by partitioning the datasets by half into the training set and validation set. The final results are obtained by averaging the performances on the validation set over five random splits. We report our results using the cumulated matching characteristic (CMC) [3], which is able to measure rank-k correct match rates.

We use a ResNet-34 model as our model feature extractor. We initialize the learning rate at 10^{-4} and divide it by a factor of 10 every 15 epochs, stopping at the maximum epoch of 40. Our network is trained with an SGD solver (weight decay of 10^{-4}) on an NVIDIA GeForce GTX 1080 GPU. Contrary to the work of [33], we employ a nonconventional bin distribution of size 6×3 for ROI Pooling since bounding boxes for people are typically tall and narrow.

B. Ablation Studies

We present our results alongside a number of ablation studies to provide a wide perspective of the performance of our proposed approach together with the impact of various components in the framework.

1) Results With Features of Different Granularities: In order to evaluate the effectiveness of our multigrained group Re-ID framework, we compare eight methods of different granularities including some variations: 1) only using global features [15] of the entire group (*Global*); 2) only using features of individual people (*Fine*); 3) using features of individual people and two-people subgroup (*Fine+Medium*); 4) using features of individual, two-people, and three-people subgroups (*Fine+Medium+Coarse*); 5) using our multigrained framework, but assigning equal importance weights for all people/people subgroups, that is, set all to 1 (*Proposed-equal weights*); 6) using our multigrained framework, but omitting

TABLE II
ABLATION STUDY RESULTS OF REPRESENTATIONS OF VARIOUS GRANULARITIES ON THE ROAD GROUP DATASET

Rank (hand-crafted)	1	5	10	15	20
Global	15.8	31.6	43.0	48.6	54.8
Fine	62.0	82.2	89.6	95.1	96.5
Fine+Medium	66.7	87.2	93.3	96.0	96.8
Fine+Medium+Coarse	71.1	89.4	94.1	97.0	97.3
Proposed-equal weights	55.8	78.0	88.1	92.1	93.6
Proposed-no spatial	69.6	88.6	94.0	96.2	96.5
Proposed-auto	72.3	90.6	94.1	97.1	97.5
Proposed-GT	**76.0**	**91.8**	**95.3**	**97.2**	**98.0**
Rank (deep convolution)	1	5	10	15	20
Global	32.1	67.9	77.8	84.0	86.4
Fine	69.1	88.9	92.3	93.8	95.1
Fine+Medium	69.1	90.1	95.1	96.3	96.3
Fine+Medium+Coarse	72.8	93.8	95.1	96.3	96.8
Proposed-equal weights	72.4	90.1	92.6	96.3	97.5
Proposed-no spatial	70.4	90.1	91.3	92.6	96.3
Proposed-auto	80.2	93.8	96.3	97.5	97.5
Proposed-GT	**82.4**	**95.1**	**96.3**	**97.5**	**98.0**

the spatial relation features in the multigrained representation (see Section III, *Proposed-no spatial*); 7) using our multigrained framework, but using the ground-truth pedestrian detection results (*Proposed-GT*); and 8) using our multigrained framework with an automatic pedestrian detection method of [43] to identify individual people in groups (*Proposed-auto*).

Table II shows the CMC results of group Re-ID on the *Road Group* dataset, measuring the correct match rates for different Re-ID ranks. The upper part lists the results based on handcrafted features (see Section III-A) while the lower part lists the results based on deep convolutional features (see Section III-B).

Fig. 7 shows some group-wise matching results under different methods. We make the following observations.

1) The *Global* method achieves poor results. This implies that simply using the entire group image cannot effectively handle the intricate variations that are present in group Re-ID. Comparatively, the *Fine* method has obviously better performance by extracting and matching individual people to handle the challenges of group dynamics. However, its performance is still hindered by the interference of pedestrian misdetections or mismatches [see Fig. 7(a)]. These problems are reduced more effectively in the *Fine+Medium* and *Fine+Medium+Coarse* methods, both of which contain subgroup-level information that captures underlying group dynamics. Our proposed framework (*Proposed-GT* and *Proposed-auto*), which includes all levels of granularity, can achieve the best performance.
2) The *Proposed-equal weights* method has obviously poorer results than its counterpart with importance weights (*Proposed-GT* and *Proposed-auto*). This clearly indicates that: a) assigning importance weights to different individuals/people subgroups is significant in guaranteeing group Re-ID performances and b) our proposed importance evaluation scheme is effective in finding proper importance weights for all levels of granularity, such that reliable and discriminative individuals/people subgroups are highlighted, resulting in better matching results. For instance, in Fig. 7(c) and (d), due to large layout change between groups, the *Proposed-equal weights* scheme is unable to assign a high score on the pairs, while the *Proposed-auto* scheme allows salient objects to be given greater importance, hence resulting in a higher matching score.
3) The *Proposed-no spatial* method achieves relatively satisfactory results. This indicates that even when spatial relation features are not encoded, our approach can generally still obtain reliable performances, propelled by multigrained information and importance weights. In the case of deep convolution features, we observe a relatively larger performance drop when spatial features are not used (about 10% for rank-1), which indicates that the choice of spatial features extracted from our CNN is crucial and can boost the group Re-ID accuracy to a large extent.
4) The *Proposed-auto* method has almost similar results as the *Proposed-GT* method, only marginally lower in most cases. The close performances of these two methods indicate that our multigrained group Re-ID framework has the ability to handle matching errors caused by pedestrian misdetections. For example, in Fig. 7(a), the left group in camera *A* is incorrectly matched with the blue rectangle in camera *B* which detected a parked motorcycle. However, by integrating multigrained information, we can successfully avoid this mismatch by considering subgroup correlation at higher-level granularities [see Fig. 7(b)]. Note that the iterative procedure refines the importance weights even at fine granularity, subsequently resulting in a correct individual match.
5) By comparing the results as a whole, we find a similar trend with respect to the combination of granularities, which affirms the good scalability and extensibility of our group Re-ID framework to accommodate different types of features. Further, we find that when deep convolutional features are used, the performance of our framework is usually better than the handcrafted counterpart on identical granularity settings.

2) Results With Different Detection Recalls: In Section VI-B1, we demonstrate that the matching accuracy is competitive when a high-quality pedestrian detector is used. Following this observation, we further investigate the effect of detection recalls on our final matching accuracy. We conduct this experiment by altering the output confidence threshold of our detector [43] to obtain detection results at different recall rates. We then perform group Re-ID based on the detected objects and their respective bounding boxes. In Fig. 8, we compare the Rank-1 CMC scores for handcrafted and deep convolutional features against detection recall rates.

From Fig. 8, we find the final matching results are relatively robust against the quality of detectors that we adopt. This is evident as the drop in Rank-1 CMC score is less than 3% for every corresponding 5% decrease (approximately) in the recall rate. This observation further demonstrates that our multigrained matching framework is robust to the detection

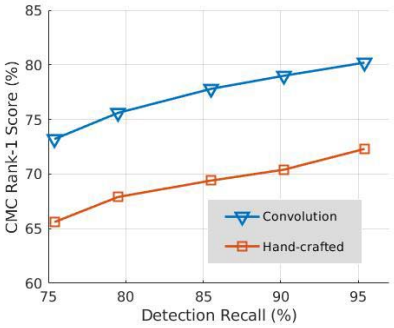


Fig. 8. Rank-1 CMC scores via different detection recall rates using hand-crafted features (orange) and convolution features (blue). (Best viewed in color.)

quality and could still be helpful for the group Re-ID problem even when prior knowledge of individuals is incomplete.

3) Results With Different Matching Constraints: In this paper, we further evaluate the effectiveness of our multiorder matching process when matching constraints are varied. Five methods are compared: 1) only single-level matching, that is, the first-order potential in (23) (*Single*); 2) discard the interorder potential term, that is, setting all $\mathcal{P}_{rl}(\mathcal{C})$ terms in (22) to 0 (*No inter*); 3) discard the unmatched term [i.e., the second term in (31)] when calculating matching scores (*No dis*); 4) use hyperedge matching [39], which integrates multigrained information by constructing a multiorder similarity function (*Hyp-E*); and 5) our proposed matching process (*Proposed-auto*).

Table III shows the CMC matching scores for the group Re-ID task using multiorder matching with different matching constraints and criterion. We report results for both handcrafted features and convolutional features on the *Road Group* dataset. From Table III, we can draw the following conclusions.

1) In comparison with the *Single* method, the higher-order potentials of the *Proposed-auto* method clearly play an important role in improving the group Re-ID score by a great measure. These potentials are essential to handle matching of multiple-person subgroups to complement the use of multigrained object representation.
2) The *Proposed-auto* method obtained better results than the *No inter* method. This comparison indicates that the interorder potential term is useful to properly capture correlations between different levels of granularity.
3) The *Proposed-auto* method performed significantly better than the *No dis* method. This demonstrates the importance of including the information of unmatched objects [see (31)] in the matching process. Results on both features show that this information has far more of an impact than that of the interorder potential term [see method (2)].
4) The *Proposed-auto* method also has better matching accuracy than the *Hyp-E* method. This demonstrates that our multiorder matching process can make better use of the multigrained information in groups during matching.

4) Results With Different Feature Combinations: We also investigate the effectiveness of using convolutional features from our multitask CNN for different parts of the extracted features [see Fig. 3(a)]. We evaluate the CMC score for all three datasets on the following combinations of features.

1) Use handcrafted features to describe both appearance and spatial relation for objects of all granularity (*Hand-crafted*).
2) Use convolutional features only to represent the global image and keep other features as handcrafted (*Global-conv*).
3) Represent appearance of global and local objects with convolutional features and keep spatial relation features handcrafted (*Appearance-conv*).
4) Convolutional features for all parts including the spatial relation features (*Full-conv*).

We report the matching accuracy with respect to different ranks in Table IV. From these results, we can observe the following:

1) Benefits are limited when only the global handcrafted feature is replaced with one that is deep convolutional. This shows that convolutional features are not sufficiently discriminative when applied to the entire group image.
2) There is a leap in improvement from the *Appearance-conv* to *Full-conv* method, which indicates that using deep representations to encode spatial relations between individuals is more impactful than opting for handcrafted representations.
3) Overall, the improvement brought on by the *Full-conv* method over the *Hand-crafted* method is more prominent on the *DukeMTMC Group* and *Road Group* datasets than on *i-LIDS MCTS*. This is indicative of the robustness of our proposed deep convolutional features in more crowded scenarios. However, the handcrafted feature is still valuable since it can be applied to any scenario without an additional training process, especially when the data are limited and/or there are insufficient means to train a CNN feature extractor.

TABLE III
ABLATION STUDY RESULTS OF VARIOUS MATCHING ORDER AND CRITERION ON THE ROAD GROUP DATASET

Rank (hand-crafted)	1	5	10	15	20
Single	62.0	82.2	89.6	95.1	96.5
No inter	70.1	88.8	94.1	96.3	97.5
No dis	65.8	88.8	93.8	96.3	96.3
Hyp-E [39]	55.1	77.8	87.6	88.9	95.1
Proposed-auto	**72.3**	**90.6**	**94.1**	**97.1**	**97.5**
Rank (deep convolution)	1	5	10	15	20
Single	69.1	88.9	92.3	93.8	95.1
No inter	74.1	92.6	96.3	97.5	97.5
No dis	66.7	90.1	95.8	96.3	97.5
Hyp-E [39]	62.9	77.8	88.9	95.1	96.3
Proposed-auto	**80.2**	**93.8**	**96.3**	**97.5**	**97.5**

C. Results on Single-Person Re-ID

Although our approach is designed for the group Re-ID task, the intermediate result of fine-grained mapping $\mathcal{C}$ could be seen as a side product of our matching process. To further investigate how multiorder constraints and priors of group

TABLE IV
CMC RESULTS FOR GROUP RE-ID ON DIFFERENT DATASETS BASED ON VARIOUS FEATURE COMBINATIONS

Rank / Features	i-LIDS MCTS					DukeMTMC Group					Road Group				
	1	5	10	15	20	1	5	10	15	20	1	5	10	15	20
Hand-craft	37.9	64.5	79.4	91.5	93.8	47.4	68.1	77.3	83.6	84.7	72.3	90.6	94.1	97.1	97.5
Global-conv	31.3	56.9	73.1	85.6	91.2	42.1	67.8	79.0	84.6	86.2	75.3	93.8	96.3	96.3	97.5
Appearance-conv	35.6	66.2	80.6	87.9	95.0	44.9	73.1	82.7	89.9	93.3	76.8	92.3	95.1	97.5	97.5
Full-conv	**38.8**	**65.7**	**82.5**	**93.8**	**98.8**	**48.4**	**75.2**	**89.9**	**93.3**	**94.4**	**80.2**	**93.8**	**96.3**	**97.5**	**97.5**

TABLE V
RANK-1 RESULTS (R1) OF DIFFERENT MATCHING SCHEMES FOR SINGLE-PERSON RE-ID ON THE ROAD GROUP DATASET

	Methods	**Rank-1 Score**
Without Group	TriNet [12]	37.2
	AlignReID [26]	32.8
	Single-match (hand-crafted)	26.5
	Single-match (deep conv)	21.0
With Group	Intra-group (hand-crafted)	67.1
	Intra-group (deep conv)	70.3
	Proposed-auto (hand-crafted)	71.4
	Proposed-auto (deep conv)	**73.5**

pairs affect the accuracy of individual matching, we conduct an extra experiment on single-person Re-ID.

We compare two state-of-the-art single Re-ID methods, that is, TriNet [12] and AlignReID [26], against four variants of our matching scheme on the *Road Group* dataset.

1) Given a person from a probe image, we find the nearest person from among all individuals in the gallery groups based on the Euclidean distance in feature space (*Single-match*).
2) Given a probe image, we first find its matched group in the gallery, and for each individual in the probe group image, we find the nearest person from among the individuals in the matched gallery group based on the Euclidean distance in feature space (*Intragroup*).
3) We first obtain the matched pairs $\mathcal{C}$ between groups by solving the group-wise multiorder matching problem, if the matched group is exactly the ground truth, we take $\mathcal{C}$ as matching results for individual objects in the probe image; otherwise, we regard all people in probe groups as unmatched (*Proposed-auto*). We conduct these experiments on both handcrafted and deep convolutional features.

We report the Rank-1 CMC score for single-person Re-ID on the Road Group dataset in Table V and further visualize some sample results of different schemes in Figs. 9 and 10. Table V is split into two parts. The upper part lists results from methods without prior for groups, while the lower part lists results with group constraint. From these results, we observe the following:

1) From Table V, we observe that a simple person-wise matching strategy without prior groups performs rather poorly compared with other approaches. This indicates person Re-ID using only person-wise descriptors may be ill-suited for such group scenarios since the search space is likely too large with limited samples per person. Without group priors, it is common to yield matched individuals from other groups who are similar in appearance [illustrated by the blue arrows in Fig. 9(a) and (b)].

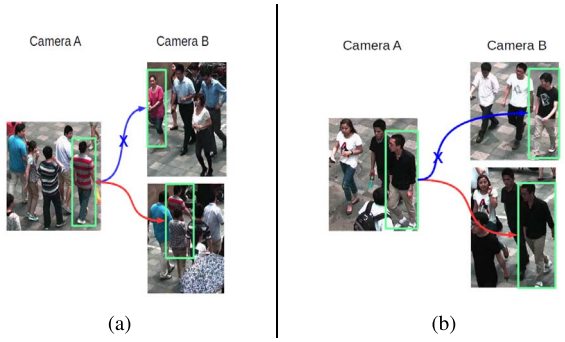


Fig. 9. Examples of incorrect individual matching (blue arrows) under *single-match* scheme and corresponding results under *proposed* scheme (red arrows). (Best viewed in color.)

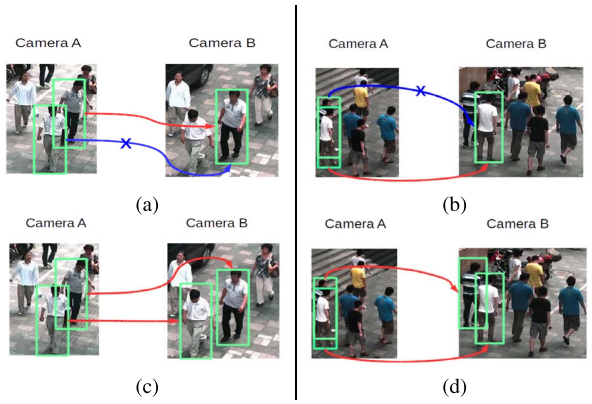


Fig. 10. Examples of individual matching results. (a) and (b) Matching results under the *intragroup* scheme. (c) and (d) Corresponding results under the *proposed* scheme. The blue arrow indicates incorrect matches while red arrows denote the correct ones. (Best viewed in color.)

2) Although the *Intragroup* method performs better than person-wise matching, it is still worse than the *proposed-auto* scheme. Even by limiting the search space of the matched group, there still exists interference that results in incorrect individual matching For example, in Fig. 10, the matched results in Fig. 10(a) show confusion with an individual with highly similar clothes; in Fig. 10(b), there exists an occlusion in an overcrowded area. Both of these factors result in failures in the *intragroup* method, while the *proposed-auto* scheme could handle these issues better since the constraint of the multiorder potential requires the system to consider optimal matching not only between individuals but also between subgroups of multiple people.

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

Table VI summarizes the group Re-ID performances on various datasets, comparing our proposed approach against

TABLE VI
CMC RESULTS FOR GROUP RE-ID ON THE THREE DATASETS

Method \ Rank	i-LIDS MCTS					DukeMTMC Group					Road Group				
	1	5	10	15	20	1	5	10	15	20	1	5	10	15	20
Saliency [6]	26.1	48.5	67.5	80.3	89.9	13.9	33.3	51.5	59.8	66.3	48.6	73.6	82.2	86.2	90.1
Mirror+KMFA [15]	28.3	58.4	69.8	80.5	90.6	11.0	31.5	49.7	62.9	70.8	25.7	49.9	59.5	66.9	72.1
CRRRO-BRO [4]	23.3	54.0	69.8	76.7	82.7	9.9	26.1	40.2	54.2	64.9	17.8	34.6	48.1	57.5	62.2
Covariance [5]	26.5	52.5	66.0	80.0	90.9	21.3	43.6	60.4	70.3	78.2	38.0	61.0	73.1	79.0	82.5
PREF [7]	30.6	55.3	67.0	82.0	92.6	22.3	44.3	58.5	67.4	74.4	43.0	68.7	77.9	82.2	85.2
BSC+CM [8]	32.0	59.1	72.3	82.4	93.1	23.1	44.3	56.4	64.3	70.4	58.6	80.6	87.4	90.4	92.1
TriNet (local) [12]	25.0	53.2	65.6	78.2	84.4	37.1	57.3	66.3	71.9	79.9	67.8	87.7	88.9	93.8	96.3
AlignReID (local) [26]	28.1	56.3	68.8	75	87.5	32.6	51.2	59.6	66.3	71.2	69.8	87.4	94.1	94.1	96.3
TriNet (global) [12]	33.6	55.0	69.4	77.5	86.9	23.6	42.7	60.7	69.7	74.2	34.6	65.4	82.7	82.4	90.2
AlignReID (global) [26]	34.4	62.5	75.0	84.3	93.7	18.0	43.8	55.1	66.3	77.5	39.5	55.6	70.4	77.8	85.9
Proposed-auto (hand)	37.9	64.5	79.4	91.5	93.8	47.4	68.1	77.3	83.6	87.4	72.3	90.6	94.1	97.1	97.5
Proposed-auto (conv)	**38.8**	**65.7**	**82.5**	**93.8**	**98.8**	**48.4**	**75.2**	**89.9**	**93.3**	**94.4**	**80.2**	**93.8**	**96.3**	**97.5**	**97.5**

state-of-the-art group Re-ID methods: *CRRRO-BRO* [4], *Covariance* [5], *PREF* [7], and *BSC+CM* [8]. For clarity, we denote the features used in our method using the suffix *hand* for handcrafted features and *conv* for the convolutional features derived from our multitask deep CNN.

For further benchmarking, we also include the results of the state-of-the-art methods designed for single-person Re-ID. Among them are methods that utilize patch saliency (*Saliency* [6]) or a KMFA(R_{χ^2}) distance metric to calculate image-wise similarity (*Mirror+KMFA* [15]). We also compare with two deep metric learning-based methods: 1) TriNet [12], a combination of CNN and triplet loss and 2) AlignReID [26], a CNN-based method which simultaneously learns global and local distances between sample images. Since these two methods are originally designed for person Re-ID, we design two variants to extend them for the group Re-ID scenario. One variant extracts features of individuals, under their respective deep frameworks [12], [26], and proceeds to apply the Kuhn–Munkres algorithm for bipartite matching between individuals in two groups. Finally, the similarity between two groups is computed as the inverse of the summation of feature distances between matched pairs. We denote this variant with a suffix *local*, named after the nature of this method. The other variant directly takes the group image as the input of the algorithm in [12] and [26] and calculates the group similarity according to the Euclidean distance between output features. We denote this variant with a suffix *global*, since it considers the entire image. From Table VI, we can observe the following:

1) Our approach (handcrafted or deep convolutional features) has better results than the other competing methods, on all three evaluated datasets. This demonstrates the resounding consistency and effectiveness of our approach in addressing the group Re-ID problem.
2) Group Re-ID methods that used global features (*CRRRO-BRO* [4], *Covariance*[5], and *PREF* [7]) achieve less satisfactory results. This indicates that utilizing only global features is clearly inadequate at handling the diverse range of challenges in group Re-ID.
3) Although the *BSC+CM* method obtained better results than that of global feature-based methods by introducing fine-grained objects (i.e., patches) to handle group dynamics, its performance is still evidently lower than our approaches. This implies the usefulness of including information from multiple granularities.
4) Our approach is also obviously superior to the conventional methods for single-person Re-ID (*Saliency*, *Mirror+KMFA*). This indicates that the task of re-identifying each individual in a group-wise setting is a rather limited solution that is likely to fail in challenging group Re-ID scenarios.
5) Deep metric-learning methods perform poorer than our approach since they only resort to fine-grained representation (between person objects) while ignoring more complex patterns that occur at the medium and coarse subgroup levels. Interestingly, this comparison also shows that these single Re-ID methods perform relatively better on groups with more individuals (e.g., *Road Group*).
6) The improvement of our approach is more obvious on datasets with larger group layouts and group member changes (*DukeMTMC Group* and *Road Group*). This demonstrates that our approach is capable of handling the dynamic changes that naturally occur in the group membership. On the other hand, the improvement on the *i-LDS MCTS* dataset is less obvious. This is the result of a limited volume of people in this dataset, since for such scenarios, representations based on multiple levels of granularity are less discriminative than when applied to crowded scenes; purely global descriptors appear to be sufficiently competitive for characterizing such small groups.

TABLE VII
RUNNING TIME ON THE THREE DATASETS

Datasets	i-LIDS MCTS	DukeMTMC Group	Road Group
All image pairs	1.1 min	18.9 min	11.5 min
Per image pair	0.06 sec	0.14 sec	0.10 sec

E. Computational Complexity

Table VII shows the running time of our group Re-ID approach on different datasets (excluding the time consumed for object detection and feature extraction). The running test is conducted on an 8-core i7-7700@3.60 GHz CPU platform.

We list two time complexity values: 1) the running time for the entire process (*all image pairs in the dataset*) and 2) the average running time for computing the similarity of a single group image pair (*per image pair*). Table VII shows that our approach is acceptable in running time.

VII. Conclusion

This paper introduces a novel approach to address the seldom-studied problem of group Re-ID. This paper contributes broadly in these aspects: 1) a multigrained group Re-ID framework which derives feature representations for multigrained objects and iteratively evaluates their importance at different granularities to handle group dynamics; 2) a multiorder-matching process which integrates multigrained information to obtain more reliable group matching results; and 3) two independent pipelines (handcrafted and deep learning) which are capable of encoding appearance and spatial relations of multigrained objects. Overall, our extensive experiments demonstrate the viability of our approaches. We also release our group Re-ID datasets involving realistic challenges to spur future works toward this direction.

References

[1] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern. Recogn.*, vol. 48, no. 10, pp. 2993–3003, 2015.

[2] S.-Z. Chen, C.-C. Guo, and J.-H. Lai, "Deep ranking for person re-identification via joint representation learning," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2353–2367, May 2016.

[3] W. Lin *et al.*, "Learning correspondence structures for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2438–2453, May 2017.

[4] W.-S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *Proc. BMVC*, 2009, pp. 1–11.

[5] Y. Cai, V. Takala, and M. Pietikäinen, "Matching groups of people by covariance descriptor," in *Proc. ICPR*, 2010, pp. 2744–2747.

[6] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 356–370, Feb. 2017.

[7] G. Lisanti, N. Martinel, A. Del Bimbo, and G. L. Foresti, "Group re-identification via unsupervised transfer of sparse features encoding," in *Proc. ICCV*, 2017, pp. 2468–2744.

[8] F. Zhu, Q. Chu, and N. Yu, "Consistent matching based on boosted salience channels for group re-identification," in *Proc. ICIP*, 2016, pp. 4279–4283.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[11] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. CVPR*, 2017, pp. 1320–1329.

[12] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arxiv pre-print: 1703.07737*, 2017.

[13] W. Zhang, B. Ma, K. Liu, and R. Huang, "Video-based pedestrian re-identification by adaptive spatio-temporal appearance model," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 2042–2054, Apr. 2017.

[14] B. Ma, Y. Su, and F. Jurie, "Covariance descriptor based on bio-inspired features for person re-identification and face verification," *Image Vis. Comput.*, vol. 32, nos. 6–7, pp. 379–390, 2014.

[15] Y.-C. Chen, W.-S. Zheng, and J. Lai, "Mirror representation for modeling view-specific transform in person re-identification," in *Proc. IJCAI*, 2015, pp. 3402–3408.

[16] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. CVPR*, 2016, pp. 1239–1248.

[17] S. Tan, F. Zheng, L. Liu, J. Han, and L. Shao, "Dense invariant feature based support vector ranking for cross-camera person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 2, pp. 356–363, Apr. 2016.

[18] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "Learning deep neural networks for vehicle Re-ID with visual-spatio-temporal path proposals," in *Proc. ICCV*, 2017, pp. 1918–1927.

[19] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury, "Consistent re-identification in a camera network," in *Proc. ECCV*, 2014, pp. 330–345.

[20] L. Zheng *et al.*, "MARS: A video benchmark for large-scale person re-identification," in *Proc. ECCV*, 2016, pp. 868–884.

[21] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proc. CVPR*, 2017, pp. 3346–3355.

[22] J. Xiao, Y. Xie, T. Tillo, K. Huang, Y. Wei, and J. Feng, "IAN: The individual aggregation network for person search," *CoRR*, vol. abs/1705.05552, pp. 332–340, Mar. 2017.

[23] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. CVPR*, Columbus, OH, USA, 2014, pp. 152–159.

[24] W. Chen, X. Chen, J. Zhang, and K. Huang, "A multi-task deep network for person re-identification," in *Proc. AAAI*, 2016, pp. 3988–3994.

[25] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," in *Proc. IJCAI*, 2017, pp. 2194–2200.

[26] X. Zhang *et al.*, "AlignedReID: Surpassing human-level performance in person re-identification," *arxiv pre-print: 1711.08184*, 2017.

[27] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, 2015, pp. 815–823.

[28] W. Liao, M. Y. Yang, N. Zhan, and B. Rosenhahn, "Triplet-based deep similarity learning for person re-identification," in *Proc. ICCV Workshop*, 2017, pp. 385–393.

[29] A. Bialkowski, P. Lucey, X. Wei, and S. Sridharan, "Person re-identification using group information," in *Proc. DICTA*, 2013, pp. 1–6.

[30] S. M. Assari, H. Idrees, and M. Shah, "Human re-identification in crowd videos using personal, social and environmental constraints," in *Proc. ECCV*, 2016, pp. 119–136.

[31] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[32] M. Cho, K. Alahari, and J. Ponce, "Learning graphs to match," in *Proc. ICCV*, Sydney, NSW, Australia, 2013, pp. 25–32.

[33] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2015.

[34] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. CVPR*, 2016, pp. 1335–1344.

[35] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. SIGMOD*, 2000, pp. 93–104.

[36] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. CVPR*, 2013, pp. 3586–3593.

[37] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, 2000.

[38] J. Yan, C. Li, Y. Li, and G. Cao, "Adaptive discrete hypergraph matching," *IEEE Trans. Cybern.*, vol. 48, no. 2, pp. 765–779, Feb. 2018.

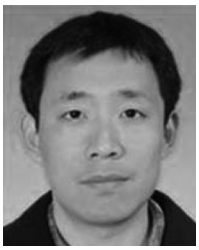
[39] J. Lee, M. Cho, and K. M. Lee, "Hyper-graph matching via reweighted random walks," in *Proc. CVPR*, 2011, pp. 1633–1640.

[40] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. ECCV Workshop*, 2016, pp. 17–35.

[41] F. Solera, S. Calderara, and R. Cucchiara, "Socially constrained structural learning for groups detection in crowd," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 87–99, May 2016.

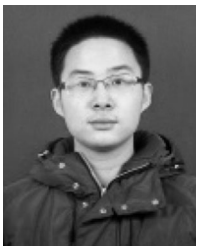
[42] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. ECCV*, 2008, pp. 262–275.

[43] L. Liu, W. Lin, L. Wu, Y. Yu, and M. Y. Yang, "Unsupervised deep domain adaptation for pedestrian detection," in *Proc. ECCV Workshop*, 2016, pp. 676–691.



Weiyao Lin received the B.E. and M.E. degrees in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2003 and 2005, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Washington, Seattle, WA, USA, in 2010.

He is currently a Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University. His current research interests include image/video processing, video surveillance, and computer vision.



Yuxi Li received the B.E. degree in electrical engineering and information from Shanghai Jiao Tong University, Shanghai, China, in 2018, where he is currently pursuing the master's degree in electrical engineering.

His current research interests include computer vision and machine learning.

Hao Xiao received the B.S. degree in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2017. He is currently pursuing the master's degree in electrical and computer engineering from the University of Washington, Seattle, WA, USA.

His current research interests include computer vision, machine learning, and autonomous driving.

John See received the B.Eng., M.Eng.Sc., and Ph.D. degrees in electrical and computer engineering from Multimedia University, Cyberjaya, Malaysia.

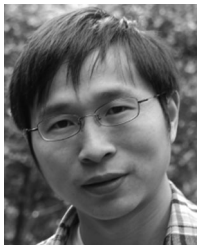
He is currently a Senior Lecturer and the Head of the Visual Processing Laboratory, Faculty of Computing and Informatics, Multimedia University. His current research interests include computer vision, pattern recognition, video processing, and affective computing.

Junni Zou received the M.S. and Ph.D. degrees in communication and information systems from Shanghai University, Shanghai, China, in 2004 and 2006, respectively.

She is currently a Full Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai. Her current research interests include video analysis, and multimedia processing and communication.

Hongkai Xiong received the Ph.D. degree in communication and information systems from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2003.

Since 2003, he has been with the Department of Electronic Engineering, SJTU, where he is currently a Full Professor. His current research interests include signal processing, computer vision, and machine learning.



Jingdong Wang received the M.E. and B.E. degrees in automation from Tsinghua University, Beijing, China, in 2001 and 2004, respectively, and the Ph.D. degree in computer science from the Hong Kong University of Science and Technology, Hong Kong.

He is a Senior Researcher with the Visual Computing Group, Microsoft Research Asia, Beijing. His current research interests include efficient CNN architecture design, person re-identification, and salient object detection.

Tao Mei (F'19) received the B.E. and Ph.D. degrees in electrical and computer engineering from the University of Science and Technology of China, Hefei, China, in 2001 and 2006, respectively.

He was a Senior Research Manager with Microsoft Research Asia, Beijing, China. In 2018, he joined JD.COM, Beijing, where he is the Deputy Managing Director of AI Research and serves as the Director of the Computer Vision and Multimedia Laboratory. He has authored or coauthored over 200 publications (with 11 best paper awards).

Dr. Mei was elected as a fellow of IAPR in 2016, a Distinguished Scientist of ACM in 2016, and a Distinguished Industry Speaker of the IEEE Signal Processing Society in 2017, for his contributions to large-scale multimedia analysis and applications.