DotSCN: Group Re-Identification via Domain-Transferred Single and Couple Representation Learning

Ziling Huang, *Student Member, IEEE*, Zheng Wang, *Member, IEEE*, Chung-Chi Tsai, *Member, IEEE*, Shin'ichi Satoh, *Member, IEEE*, and Chia-Wen Lin, *Fellow, IEEE*

Abstract—Group re-identification (G-ReID) is an important yet less-studied task. Its challenges not only lie in appearance changes of individuals, but also involve group layout and membership changes. To address these issues, the key task of G-ReID is to learn group representations robust to such changes. Nevertheless, unlike ReID tasks, there still lacks comprehensive publicly available G-ReID datasets, making it difficult to learn effective representations using deep learning models. In this article, we propose a Domain-Transferred Single and Couple Representation Learning Network (DotSCN). Its merits are two aspects: 1) Owing to the lack of labelled training samples for G-ReID, existing G-ReID methods mainly rely on unsatisfactory hand-crafted features. To gain the power of deep learning models in representation learning, we first treat a group as a collection of multiple individuals and propose transferring the representation of individuals learned from an existing labeled ReID dataset to a target G-ReID domain without a suitable training dataset. 2) Taking into account the neighborhood relationship in a group, we further propose learning a novel couple representation between two group members, that achieves better discriminative power in G-ReID tasks. In addition, we propose a weight learning method to adaptively fuse the domain-transferred individual and couple representations based on an L-shape prior. Extensive experimental results demonstrate the effectiveness of our approach that significantly outperforms state-of-the-art methods by 11.7% CMC-1 on the Road Group dataset and by 39.0% CMC-1 on the DukeMCMT dataset.

Index Terms—Group re-identification, domain transfer, couple representation, video surveillance, deep learning.

I. Introduction

DUE to the fast-growing applications in security and surveillance, person/vehicle re-identification (ReID) has been drawing much attention [1]–[9]. While existing research works mainly focused on re-identifying individuals, searching out a group of multiple persons simultaneously was relatively rarely studied. In many practical applications identifying a group and then tracking and analyzing the group's behaviors and activities is of fundamental importance. Hence, re-identifying a group of persons (Group ReID or G-ReID) across cameras in different environments is getting more and more important. The G-ReID problem is different from the ReID problem from two perspectives. From the research perspective, G-ReID poses new challenging research problems (will be elaborated in the following) when the ReID target becomes a group of persons. These problems cannot be effectively addressed by existing individual ReID methods as evidenced from the unsatisfactory performances of existing Re-ID methods on G-ReID, and thus call for the development of novel group representations and solutions. From the application perspective, G-ReID is a powerful supplement to individual ReID. For example, criminal cases are often conducted by a certain group of persons rather than an individual. Individual ReID techniques, however, often cannot identify a whole group of suspects in such an application scenario. The technical challenges and practical values of G-ReID have attracted significant amounts of research efforts along this direction.

Different from ReID, G-ReID aims at associating a certain group across different cameras. Besides the challenge of appearance changes for individuals, such as low-resolution, pose variation, illumination variation, and blurred vision, G-ReID raises novel and unique challenges. 1) *Group layout change*. People in a group often change their locations under different camera views, as Fig. 1 shows. 2) *Group membership change*. Also as illustrated in Fig. 1, people will dynamically leave or join a group, thereby making the number of persons in
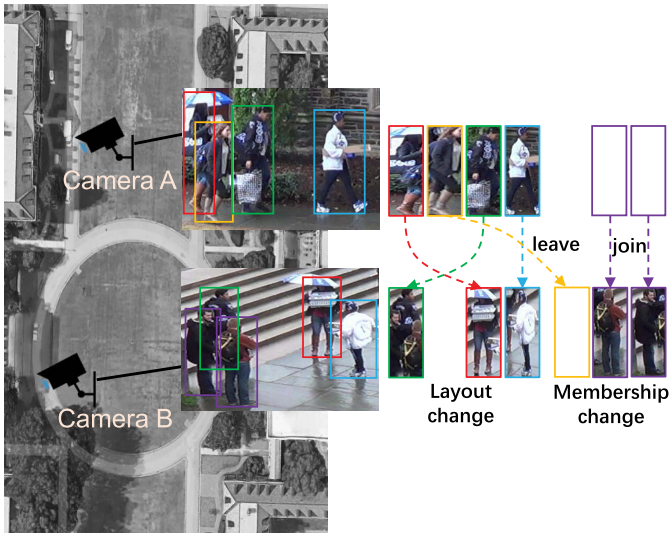
Fig. 1. **Illustration of the challenges for G-ReID.** Besides the challenge of the appearance change, G-ReID brings in the challenges of group layout and membership changes.

TABLE I

COMPARISON OF THE CHALLENGES IN REID AND G-REID

Challenge	ReID	G-ReID
Appearance Change	✓	✓
Layout Change	×	✓
Membership Change	×	✓
Training Set	Abundant	No

the group change over time. As a result, it is usually not a good choice to treat the group as a whole to extract its global/semi-global features as [10] did, because the temporal dynamics in a group's layout and membership will alter the visual content of the group. 3) *Lack of labelled training data*. In addition, G-ReID is a new task and lacks annotated image samples with group IDs, *i.e.*, the number of group images is too few to learn robust group representations. We list a comparison of ReID and G-ReID as shown in Table I.

Based on the considerations above, We focus on the two most challenging problems in G-ReID: 1) the problem of scarce training data, and 2) the problem of layout and membership changes and propose the following new representation learning schemes:

A. Domain-Transferred Single Representations

Since global features cannot well represent dynamically varying contents in a group, we consider to use single person features to perform group matching as a group image can be treated as a collection of multiple subimages of individual group members. The reason behind is that if we find corresponding persons in a target group, we will search out the target group as well. Considering that training data is difficult to acquire, [11] exploited hand-crafted features to represent individuals in a group. Nevertheless, hand-crafted representations usually cannot effectively address the appearance change

problem in G-ReID due to changes in environments and video capturing conditions. As we know, there exist rich amounts of training datasets suitable for general ReID, which motivates us to make use of existing labeled ReID samples to learn single-person representations. However, the domain gap between the ReID training datasets and target G-ReID images often cause severe performance drop. Therefore, to compensate for the domain shift, we need to find a way to transfer the model learned from existing ReID datasets to better represent new individuals in target G-ReID images. Motivated by the demonstrated success of [12], [13], we propose to transfer the image style of a ReID dataset to that of target G-ReID dataset while preserving individuals' identities. In this way, representative features of individuals in a group can be properly extracted by our transferred representation model.

B. Domain-Transferred Couple Representations

In G-ReID, we can obtain additional useful information from neighboring group members. Because of group membership changes, it is difficult to exactly determine how many persons are in a group. Regardless of the number of persons in a group, the group can be represented as multiple couple relations. For example, if a group contains three persons A, B and C, their couple relations expressed as $A - B$, $A - C$ and $B - C$ can be used as effective features for identifying a target group. As a result, we can search out a target group by finding a best-match with the group's corresponding couple relations. Based on this consideration, we propose a couple representation learning network to leverage the information of neighboring individuals. The couple representation learning can also benefit from the proposed domain-transfer technique mentioned above. Note that, in a group, all members can be represented as multiple one-to-one relations (couple representations). Triplet or batch relations can be also represented by multiple one-to-one relations. That is why we choose the couple representation, which is the simplest unit for social relationship.

Based on the above discussions, we propose a Domain-Transferred Single and Couple Representation Learning Network (DotSCN), that can offline learn effective features for representing single members and couple relations in a group. We also propose an online feature fusion method to adaptively combine the two kinds of features together to obtain better group representations. Motivated by the demonstrated success of [14], our method learns the weights for fusing the single representations and couple representations without supervision. As will be shown below, our method is easy to implement, yet effective..[1] Furthermore, the simplicity of our method makes it easy to be applied to different target domains, which is usually not the case for sophisticated methods.

Our contributions lie in the following three aspects:
- To tackle the scarce training data problem, we propose a novel method to transfer rich collections of individual ReID samples to the target domain of group ReID to enrich the training samples.

[1] Our code link: https://github.com/huangzilingcv/G-ReID

- We further propose a novel couple representation to capture the social relationship in a group to well address the problem of layout and membership changes for the first time.
- We propose an online fusion method to adaptively combine the learned single and couple representation results together for better group re-identification. Extensive experimental results confirm a performance leap from the relevant state-of-the-art techniques in the area.

Compared with its preliminary conference version [15], this article has been significantly expended in several aspects. First, we clarify the differences between ReID and G-ReID that motivates this work and provide a detailed survey of related works. Second, we fully reorganize the offline representation learning scheme and clarify the online adaptive feature fusion scheme based on an L-shape (turning-point) prior. Third, more quantitative tests are conducted in the experiment section along with insightful analyses on the quantitative evaluations. Specifically, we analyze the effectiveness of single and couple features and their fused representation, respectively, and the complexities of offline representation learning and extraction and online feature fusion. With these newly added components, the superiority of the proposed method is thoroughly validated.

The rest of this article is organized as follows. Some most relevant works are surveyed in Sec. II. Sec. III presents the proposed schemes for transferred representation learning, couple representation learning, and feature fusion. In Sec. IV, experimental results are demonstrated. Finally, conclusions are drawn in Sec. V.

II. RELATED WORK

A. Deep Learning Based ReID

Deep learning-based approaches have been extensively studied in general ReID field. For example, Li *et al.* [16] proposed a filter pairing neural network to jointly handle misalignment and geometric transforms. In order to learn features from multiple domains, Xiao *et al.* [17] utilized a domain-guided dropout algorithm to improve the feature learning procedure. Moreover, the method proposed in [18] makes full use of human part cues to alleviate pose variations and learn robust representations from both a whole image and its different local parts. Chen *et al.* [19] formulated a unified deep ranking framework that jointly maximizes the strengths of features and metrics. Zhu *et al.* [20] integrated spatial information for discriminative visual representations by partitioning a pedestrian image into horizontal parts, and proposed a part-based deep hashing network. Yao *et al.* [21] proposed a part loss network, to minimize both the empirical classification risk on training person images and the representation learning risk on unseen person images. Zheng *et al.* [22] introduced pose-invariant embedding as a pedestrian descriptor and designed a Pose-Box fusion CNN architecture. The descriptor is thus defined as a fully connected layer of the network for the retrieval task. However, these supervised learning-based works all require abundant labeled training data. Moreover, all of these works

mainly focused on individual person re-identification. None of them paid attention to G-ReID with very limited training data.

B. G-ReID

Recently, relatively fewer works have focused on G-ReID tasks [10], [11], [23]–[26], compared to general ReID tasks. Some of them mainly attempted to extract global or semi-global features. For example, Cai *et al.* [23] proposed a discriminative covariance descriptor to obtain both global and statistical features. Zheng *et al.* [24] proposed semi-global features by segmenting a group image into several regions. Since persons in a group often change their locations under different views (*i.e.*, layout-change), global and semi-global features are usually sensitive to such changes. In order to make use of individuals' features in the groups, Zhu *et al.* [25] introduced patch matching between two group photos. However, it requires prior restrictions on vertical misalignments, making it unworkable under certain circumstances. Lin *et al.* [11] leveraged multi-grain information and attempted to fully capture the characteristics of a group. This approach, however, involves too much redundant information and employs common hand-crafted features, thereby making its accuracy not satisfactory enough. In contrast to the aforementioned methods, our previous work [26] focused on handling the changes on layout and membership of group members using a graph-based group representation. Specifically, it treats a group of individuals as a graph, where each node denotes the individual feature and each edge represents the relation between a couple of individuals. Graph samples are then used to train a graph neural network (GNN) to learn a graph group representation for G-ReID.

C. Domain Transfer

Recently, Generative Adversarial Networks (GANs) have been applied to transfer image styles from a source domain to a target domain [27]–[31]. Gatys *et al.* [27] separated the image content and image style apart and recombined them afterwards, so that the style of one image can be transferred into another. Taigman *et al.* [30] proposed a domain transfer network to translate images to another domain while preserving original identities. By making use of the existing domain-transfer techniques, we are able to take advantage of the abundant ReID datasets to improve the performance of our method.

III. DOMAIN-TRANSFERRED SINGLE AND COUPLE REPRESENTATION LEARNING

In a G-ReID task, we have a probe image p containing a group of N persons. We aim at finding the corresponding group of probe image p in a set of gallery images $\mathcal{G} = \{g_t\}$, where g_t represents the t-th group image in gallery $\mathcal{G}$. Let g_t^j denote the j-th person in group image g_t. As depicted in Fig. 2, the proposed framework consists of three major parts. First, a domain transfer method is used to transfer the styles of the training ReID dataset into that of the target G-ReID images. In the offline learning process, three sets of transferred training data are respectively constructed: the single-person
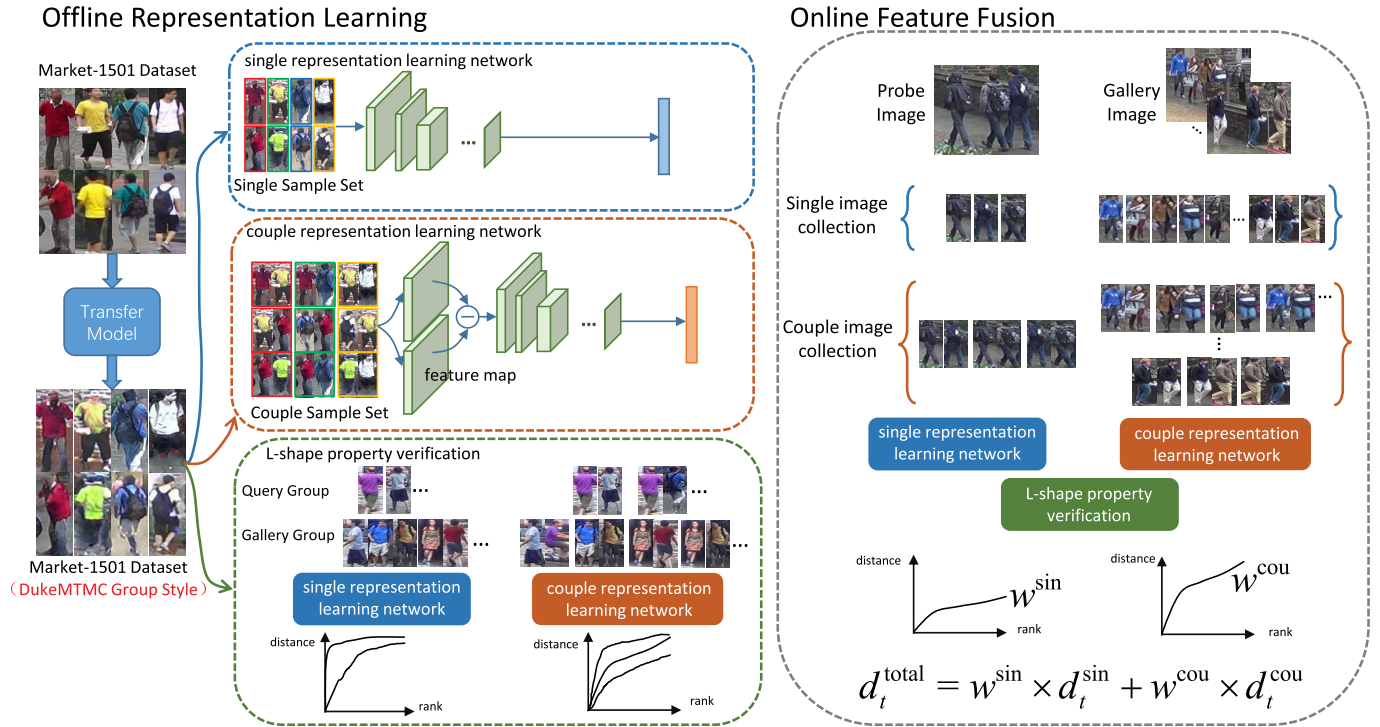
Fig. 2. **Proposed architecture.** The proposed Domain-transferred Single and Couple Network (DotSCN) consists of two parts: the offline learning process and the online feature fusion process. Here, we take DukeMTMC group dataset as the example target-domain dataset. In the offline learning process, we first transfer the style of source-domain dataset (*e.g.*, Market-1501) to that of the target one (*e.g.*, DuketMTMC). We construct three sets of transferred training data: the single-person set, couple-person set and group set. We then train DotSCN on the transferred single-person set and couple-person set. The group Set is utilized to verify the L-shape property. In the online identification process, the persons in the group set are cropped to construct the testing single-person and couple-couple sets. Through evaluating the L-shape fitness of the curves, DotSCN learns to fuse the single and couple features to obtain the final representation. Consequently, DotSCN extracts the single and couple representations of all individual members and couple combinations in a group image, and then measures the distance of the fused single and couple representation to that of gallery groups.

set, couple-person set, and group set. We train the proposed DotSCN on the single-person set and couple-person set. The group Set is utilized to verify the L-shape property as will be explained later. In the online identification process, the features extracted by the Single Representation Learning Network (SRLN) and the Couple Representation Learning Network (CRLN) are adaptively fused so that one group can be accurately identified by using the fused single and couple representations. Note that we follow the idea and setting of [11] to make the definition of a group in this article, different from that of [32]. We treat all persons in one image as a "group". If two cross-view images from different cameras share more than 60% members in common, they are considered to contain the same group.

A. Domain Transfer

Because the total number of people in a collection of G-ReID images is usually rather limited, it is difficult to train a useful network directly based on those available samples. To learn effective representations, we should make use of external information. There exists a rich collection of ReID datasets of individual persons that can help to learn good feature representations. Nevertheless, the domain gap between the existing ReID datasets and the target G-ReID images, caused by their different capturing conditions, can significantly degrade the effectiveness of representation learning. To address

this problem, given a training ReID dataset $\mathcal{S} = \{s_i\}_{i=1}^{N_s}$, we propose utilizing domain transfer to learn a mapping function $G : \mathcal{S} \rightarrow \mathcal{G}$ from the style of ReID dataset $\mathcal{S}$ to that of G-ReID dataset $\mathcal{G}$ so that the distribution of $G(\mathcal{S})$ can be indistinguishable from that of dataset $\mathcal{G}$. In our work, we exploit the CamStyle,[2] method [13] to generate the dataset $G(\mathcal{S})$ from the dataset $\mathcal{S}$.

In this way, the dataset $G(\mathcal{S})$, where $y_k^s \in G(\mathcal{S})$ denotes the k-th image of the s-th person in the dataset, can be used to train the SRLN $\mathcal{F}^{\text{sin}}$. In addition, we construct the co-occurrence relations $r^{s_1 s_2} = \{(y_{k_1}^{s_1}, y_{k_2}^{s_2})|s_1, s_2 = 1, \ldots, N_s,$ $k_1 = 1, \ldots, N_{s_1}, k_2 = 1, \ldots, N_{s_2}, s_1 \neq s_2\}$ between pairs of persons. These co-occurrence relations are used to train the CRLN $\mathcal{F}^{\text{cou}}$.

The two networks are then respectively used to extract the single and couple features of gallery images g_t and probe image p. For probe image p, let $\mathcal{F}^{\text{sin}}(p^i)$ and $\mathcal{F}^{\text{cou}}(p^{i_1 i_2})$ represent its single and couple features, respectively, where $p^{i_1 i_2} = \{(p^{i_1}, p^{i_2})|i_1, i_2 = 1, \ldots, N, i_1 \neq i_2\}$ denotes all couple pairs of probe image p. For gallery image g_t, its single and couple features are respectively represented as $\mathcal{F}^{\text{sin}}(g_t^j)$ and $\mathcal{F}^{\text{cou}}(g_t^{j_1 j_2})$, where $g_t^{j_1 j_2} = \{(g_t^{j_1}, g_t^{j_2})|j_1, j_2 = 1, \ldots, N_t, j_1 \neq j_2\}$ denotes the couple relations of a gallery image.

---

[2]We refer to the code https://github.com/zhunzhong07/CamStyle.

B. Offline Representation Learning

The offline Representation Learning Framework consists of two parts: SRLN for extracting single-person features, and CRLN for extracting joint features between two members in a group. We select ResNet-50 as the backbone CNN structure, since it is the most popular CNN network used in general ReID. Detailed settings about training the network are presented in the experiment section.

To match a group in a collection of group images, the most straightforward method is to find every corresponding people between two groups. We take advantage of an abundance of existing ReID datasets $G(\mathcal{S}) = \{y_k^s\}$ to train the SRLN $\mathcal{F}^{\text{sin}}$, for which two loss terms are used in the loss function: a cross-entropy loss for classification and a triplet loss for similarity learning.

For G-ReID, besides the personal features of individual group members, the joint features between two co-appearing members are also useful as the co-occurrences of group members have proven to be effective features for characterizing the social relationship in a group [33], [34]. As we know, no matter how many people in an group, they can always be represented as relations between every two group members. In our work, we focus on the relations between two members co-occurring in a group regardless of their spatial distance in the group image as their locations vary in a group dynamically under different camera views. Based on the above discussion, we propose to represent the couple relations between every two group members by the difference of their personal features to compactly encode the co-occurrence of the two members with only their discrepancy, since the subtraction operation removes their common features. By contrast, the co-occurrence of group members can also be represented by other operations like addition and concatenation. However, adding two person's features magnifies their common features and makes the representations less discriminative, whereas concatenation doubles the number of parameters, thereby significantly increasing the complexity of network training. Specifically, our experimental results show that adopting concatenation for training our model makes it difficult to converge.

Based on the above considerations, the goal of the CRLN $\mathcal{F}^{\text{cou}}$ is to learn effective occurrence representations of couple pairs of members in a group. To this end, after extracting the features of individual members by the SRLN, the CRLN then pairs all two-member couples in the $G(\mathcal{S})$, and for each couple, subtracts one member's features from the other's to represent their co-occurrence features $r^{s_1 s_2}$. Similar to SRLN, a loss function involving a cross-entropy loss term for classification and a triplet loss term for similarity learning is used to train the CRLN.

To utilize the cross-entropy loss to train the CRLN, we assign a unique label to each couple pair of the couple-person set based on a mapping function $\ell^{\text{cou}}(i, j)$, where i and j $(i \neq j)$ denote the two person labels in a couple-pair. The mapping function $\ell^{\text{cou}}(i, j)$ should satisfy (i) the commutative property: $\ell^{\text{cou}}(i, j) = \ell^{\text{cou}}(j, i)$; and (2) the uniqueness property: $\ell^{\text{cou}}(i_1, j_1) = \ell^{\text{cou}}(i_2, j_2)$ if and only if $\{i_1, j_1\} = \{i_2, j_2\}$. There are multiple choices of ℓ^{cou} meeting the above requirements. In our implementation, the couple-pair ID is obtained by

$$\ell^{\text{cou}}(i, j) = \frac{1}{2} \max(i, j)^2 - \frac{1}{2} \max(i, j) + \min(i, j). \quad (1)$$

Taking (1) as an example, suppose we have three identities 1, 2, 3 in a group. The couple-pair ID between samples from person 1 and person 2 is obtained as $\ell^{\text{cou}}(1, 2) = \frac{1}{2} \cdot 2^2 - \frac{1}{2} \cdot 2 - 1 = 0$. Subsequently, the couple features $r^{s_1 s_2}$ along with their unique couple-pair IDs are used to train $\mathcal{F}^{\text{cou}}$ to learn the representations of couple pairs.

C. Online Feature Fusion

Since the proposed DotSCN extracts both individuals' features and pairs' features, the two kinds of features need to be adequately fused to obtain better representations for G-ReID. Inspired by [14] which shows that given the learned representations for traditional ReID are discriminative, the rank-distance curve between a probe image and all gallery images sorted in the ascending order will exhibits an "L" shape, and vice versa. The L-shape property means that there exists a turning point at which the rank-distance curve is significantly flattened out. This property gives a clue for finding a way of fusing the two kinds of features in DotSCN. As our work is focusing on G-ReID rather than individual ReID, we need to make sure if the L-shape property still holds with G-ReID features.

To verify the aforementioned property, we use the existing Person-ReID datasets to simulate groups. Specifically, if two images contain the same group of members, they are labeled as the same group. In this step, we have the domain-transferred dataset $\mathcal{G}$ with N_s person IDs which are divided into N_{sub} subgroups equally, each containing $N_u = \lfloor \frac{N_s}{N_{sub}} \rfloor$ persons in total and being assigned with a unique group ID, where $\lfloor \cdot \rfloor$ means the floor function. Based on the consideration above, we assume that a set of $r_0 \times N_u$, where $r_0 \in [0, 1]$, members randomly picked from a group keep staying in the group, whereas the remaining $N_u - r_0 \times N_u$ members may dynamically join or leave the group. No matter how they change their locations or how many people are in the group, they share the same group ID. In this section, we randomly select one combination for each group as a probe image, denoted p, the rest of combinations are treated as gallery images, denoted g_t.

Let $\{\mathcal{F}^{\text{sin}}(p^i)\}_{i=1}^{\lfloor \frac{N_s}{N_{sub}} \rfloor}$ denote the single features of all individuals in a probe image and $\{\mathcal{F}^{\text{sin}}(g_t^j)\}_{j=1}^{N_t}$ denote the features of all individuals in the t-th gallery image g_t. The distance between probe image p and gallery image g_t is defined as

$$d_t^{\text{sin}} = \frac{1}{N} \sum_{i=1}^{N} \min\{D(\mathcal{F}^{\text{sin}}(p^i), \mathcal{F}^{\text{sin}}(g_t^j)) | j = 1, 2, \ldots, N_t\},$$

$$(2)$$

where $D(\cdot)$ denotes the distance metric (say, Euclidean distance in this work). Similarly, we also calculate the couple distance d_t^{cou} between probe image p and gallery image g_t by replacing the single features with the couple features in (2).
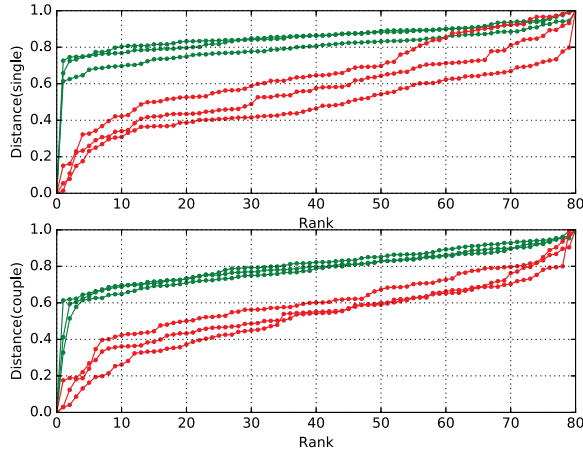
Fig. 3. **Logarithmic rank-distance curves for different probes.** Each curve indicates all the distances (sorted in the ascending order) between a probe and all the gallery groups. The top figure shows the results with single features, and the bottom one is generated using couple features. Note that the green lines indicate the representations achieving a good retrieval accuracy, whereas the red lines represent low-accuracy ones. All the high-accuracy curves (*i.e.*, the features are discriminative) are L-shaped, but the property does not hold for the low-accuracy curves.

The logarithmic rank-distance curves are shown in Fig. 3. We select $\lfloor \frac{N_s}{N_{sub}} \rfloor = 10$ and $r_0 = 0.3$ randomly. The green curves indicate the matching distances with discriminative features, while the red ones indicate the matching distances with poor features. Obviously, the green curves are L-shaped, while the red ones are not. And the green curves enclose larger areas compared with the red ones. Based on the fact above, we propose to fuse the single and couple features to obtain the final distance by applying the following weighted sum:

$$d_t^{\text{total}} = w^{\text{sin}} \times d_t^{\text{sin}} + w^{\text{cou}} \times d_t^{\text{cou}}, \tag{3}$$

where the weights are determined by the enclosed areas of the distance curves with the single and couple features, respectively:

$$w = e^{\sum_r d_t}, \tag{4}$$

where r denotes the rank (*i.e.*, the x-axis of the the curves in Fig. 3). The higher the enclosed area, the more discriminative the features, and the larger the weight.

Based on (2), the best-match gallery image can be identified by searching for that with the lowest matching distance:

$$d_p = \min\{d_t^{\text{total}} | t = 1, 2, .., N_t\}. \tag{5}$$

D. Computational Complexity Analyses

The proposed DotSCN consists of two parts. In the offline learning process, the majority of computation is consumed by learning the single and couple representation networks whose complexities depend on the number of training samples. Suppose that there are N_s person images in the training set, the computational complexity of training the single representation network is $O(N_s)$, and that of training the couple representation network is $O(N_s^2)$. Since, in practical applications, we usually extract single and couple features

for all gallery images offline in advance, their computation complexities become $O(N_t \cdot N)$ and $O(N_t \cdot N^2)$, respectively, where N_t denotes the number of persons in the gallery image set and N denotes the average number of persons in each image.

In the online fusion process, the majority of computation is spent on extracting the features of the probe image, calculating the distances between the probe image and gallery images based on the extracted features, and deriving the fusion weights. For each probe image, the time cost depends on the amount of probe and gallery features. Suppose that there are N persons in the probe image, the computation complexity will be $O(N \cdot N_t \cdot N + N^2 \cdot N_t \cdot N^2)$ or $O(N^4 \cdot N_t)$. This shows that the online computation complexity counts more on the average number of persons N in each image, implying that the scene complexity of the image dominates the complexity of our algorithm, and the number of gallery images also makes impact.

IV. EXPERIMENTAL RESULTS

A. Datasets and Experiment Setting

1) Datasets: Our method is evaluated on two public G-ReID datasets available in [11]. Some examples are shown in Fig. 4. The **DukeMTMC Group** dataset contains 177 group image pairs selected from a 8-camera-view DukeMTMC dataset [35]. The **Road Group** dataset contains 162 group pairs taken from a two-camera crowd road scene. These two datasets both include severe object occlusions and large layout & group membership changes. Following [11], half of each dataset is evaluated under the protocol in [25], and the Cumulative Matching Characteristic (CMC) metrics [11] are used for performance evaluation. Moreover, we define two cross-view groups as the same one when they have more than 60% members in common.

We also use the Market-1501 dataset [36] as the source-domain ReID dataset due to its large amount of training instances: 15,936 images for 751 individuals. In our work, we transfer the domain of Market-1501 dataset to those of **DukeMTMC Group** and **Road Group**, respectively, as illustrated in Fig. 4.

2) Setting for Domain Transfer: In our work, we transfer the domains of existing ReID datasets to that of the target G-ReID datasets (*e.g.*, **DukeMTMC Group** and **Road Group**), prior to training the representations. We use the CycleGAN [37] to transfer the domain for each target G-ReID dataset. As a result, we obtain the DukeMTMC-style Market-1501 dataset and the Road-style Market-1501 dataset. In the training stage, we resize all input images to 256×256 and use the Adam optimizer. The batch size is 10, and the learning rates are 0.0002 and 0.0001, respectively, for the Generator and the Discriminator.

3) Setting for CNN Models: In both SRLN and CRLN, the ResNet-50 with two additional fully connected layers is used as the backbone. The learning rate is set to 0.01, and the dropout rate is set to 0.5. The input images are resized

---

[2]The authors from the universities (in Taiwan and Japan) completed the experiments on the datasets
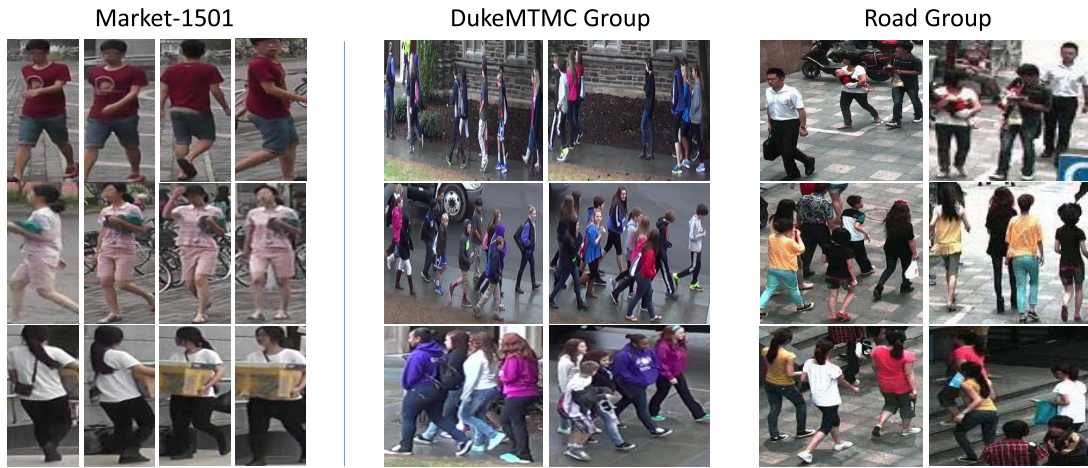
Fig. 4. **Snapshots of the utilized datasets.** From left to right, the datasets are respectively **Market-1501** (ReID), **DukeMTMC Group** and **Road Group** (G-ReID). Each row of each dataset shows a few snapshots with the same person/group ID.
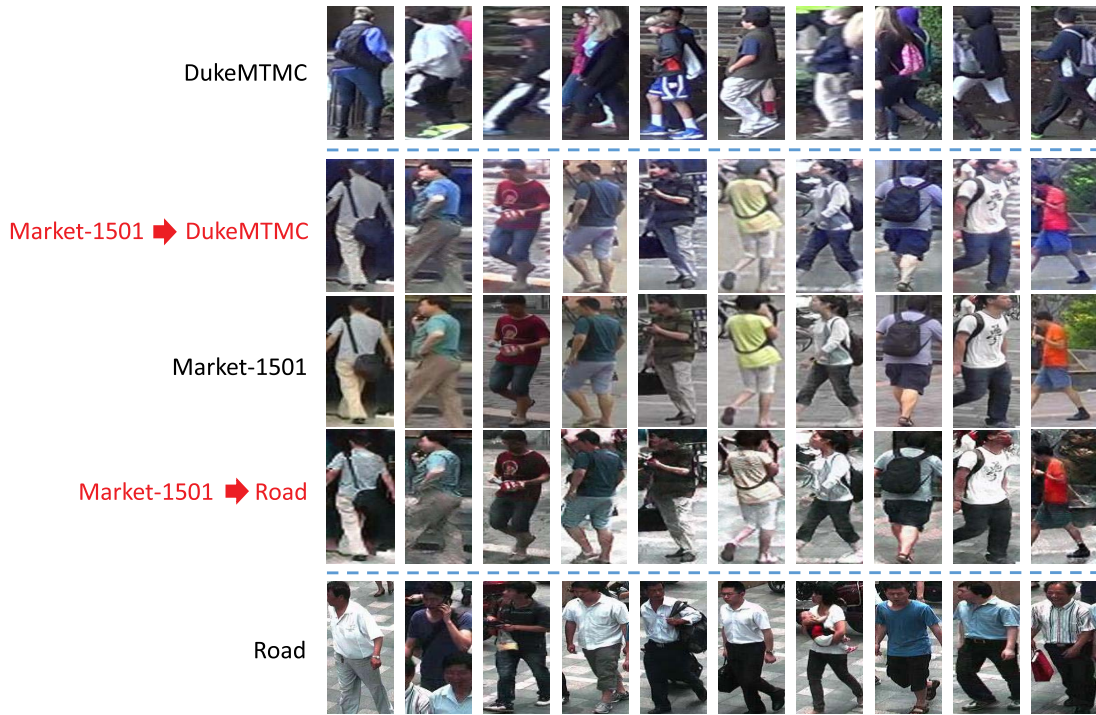

Fig. 5. **Snapshots of domain-transferred samples.** The images in the third row are from the source domain **Market-1501**. The images in the first and fifth rows are cropped respectively from the target domain **DukeMTMC Group** and **Road Group**. The second row shows the generated images with **DukeMTMC** style, and the fourth row shows the generated images with **Road** style.

to 256×128. We use the SGD solver to train the CNN model and set the batch-size to 24. To construct the training set, we pick 100 identities from the domain-transferred dataset, where each identity contains $14 - 16$ samples. If we use identity 1 and identity 2 to form paired samples for a couple class, which contain $C1$ and $C2$ samples respectively, we can generate $C1 \times C2$ couple-pairs for training.

B. Performance of Domain-Transferred Representation Learning

We conduct experiments to show the effectiveness of domain-transferred representation learning, respectively exploiting the single representations and couple representations learned on **Market-1501** with/without domain transfer, respectively. For the testing datasets for G-ReID, we use DukeMTMC/Road samples as the target-domain samples for domain transfer, as illustrated shown in Fig. 5.

Table II shows the CMC results of the single and couple representations with and without domain-transferred learning, denoted 'SCN' and 'DotSCN', respectively. From the table, we find that the deep learning features, *i.e.*, 'SCN(s)' and 'SCN(c)' beat all the hand-crafted feature based schemes in representing a group. We can also observe that the domain-transferred representation learning (*i.e.*, 'DotSCN(s)'

TABLE II

PERFORMANCE EVALUATION OF THE DOMAIN-TRANSFERRED REPRESENTATION LEARNING ON **DUKEMTMC GROUP**
AND **ROAD GROUP**. THE SUFFIX 's' INDICATES THE RESULTS WITH SINGLE REPRESENTATIONS,
AND THE SUFFIX 'c' INDICATES THE RESULTS WITH COUPLE REPRESENTATIONS

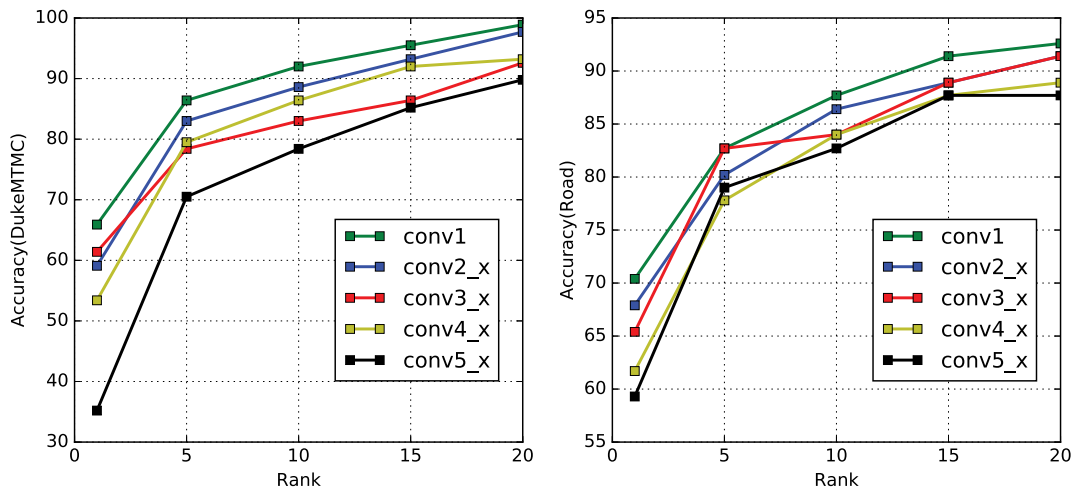| Method	DukeMTMC Group				Road Group			
	CMC-1	CMC-5	CMC-10	CMC-20	CMC-1	CMC-5	CMC-10	CMC-20
MGR	47.4	68.1	77.3	87.4	72.3	90.6	94.1	97.5
SCN(s) ResNet50	75.0	83.0	89.8	94.3	80.2	90.1	92.6	**98.8**
DotSCN(s) ResNet50	**80.7**	**89.8**	**94.3**	**96.6**	**82.7**	**93.8**	**96.3**	**98.8**
SCN(s) ResNet101	78.4	84.1	89.8	96.6	82.7	90.1	93.8	98.8
DotSCN(s) ResNet101	**88.8**	**97.7**	**98.9**	**100.0**	**83.0**	**96.6**	**97.7**	**100.0**
SCN(c) ResNet50	50.0	80.7	85.2	95.5	50.6	74.1	80.2	87.7
DotSCN(c) ResNet50	**62.5**	**83.0**	**89.8**	**95.5**	**70.4**	**79.0**	**84.0**	**90.1**
SCN(c) ResNet101	53.4	83.0	86.4	94.3	56.7	79.0	81.4	90.1
DotSCN(c) ResNet101	**77.7**	**97.7**	**97.7**	**98.9**	**74.0**	**80.2**	**85.1**	**93.8**

Fig. 6. **Effect of the selection of the subtraction layer for couple representation learning.** The curves show the G-ReID accuracy of different network designs of subtraction layers. The two figures respectively show the results (CMC-1) on **DukeMTMC Group** and **Road Group**, respectively.

and 'DotSCN(c)') outperforms the non-transferred learning ones. The single representation learning on top of the ResNet50 backbone network achieves 5.7% and 2.5% CMC-1 accuracy improvements on **DukeMTMC Group** and **Road Group**, respectively. The couple representation learning achieves 12.5% and 19.8% CMC-1 accuracy improvements on **DukeMTMC Group** and **Road Group**, respectively. Similarly, our method on top of Resnet101 also achieve significant improvements.

C. Performance of Couple Representation Learning

The CRLN is constructed by the ResNet50 network pre-trained on the domain-transferred data. We divide ResNet50 into six parts based on [38]. The difference of the feature maps of different parts can be obtained by subtraction, and is then sent to the rest of the network. The output of the fully-connected layer is taken as the couple representations. The results are shown in Fig. 6, that shows that if the feature map subtraction is placed closer to the front (conv1), we can obtain better results. That is, the CRLN performs better in

representing the relations of low-level feature maps, rather than in the high-level individual features.

Moreover, although Table II shows that the average accuracy with single representations is generally better than that with couple representations, in some cases (e.g., significant appearance variations of individuals due to changes on pose, membership and layout) the couple representations will do a better job and thereby effectively improve the discriminating power via adaptive feature fusion. For example, Fig. 7 illustrates two queries selected from the results of **DukeMTMC Group** and **Road Group**, respectively, for which couple representations are more discriminative than single representations. The first column of Fig. 7 shows the probe image. The following columns show the top three matches, where the samples highlighted by red bounding boxes indicate the negative results, while those highlighted by green bounding boxes indicate the positive results. The two examples show that, in case that the single representations lose their effectiveness, the couple representations take their responsibility and offer better predictions. Here we also compare the couple representations formed by (1) subtraction

TABLE III

PERFORMANCE EVALUATION OF DIFFERENT COUPLE REPRESENTATIONS RESPECTIVELY CONSTRUCTED BY SUBTRACTION AND ADDITION ON **DUKEMTMC GROUP** AND **ROAD GROUP**. THE ABBREVIATION 'ADD.' MEANS THE COUPLE REPRESENTATIONS CONSTRUCTED BY PERFORMING ADDITION ON FEATURE MAPS, AND 'SUB.' MEANS THE COUPLE REPRESENTATIONS CONSTRUCTED BY PERFORMING SUBTRACTION ON FEATURE MAPS

Method	DukeMTMC Group				Road Group			
	CMC-1	CMC-5	CMC-10	CMC-20	CMC-1	CMC-5	CMC-10	CMC-20
DotSCN(c) ADD.	44.3	56.8	80.7	86.4	48.1	70.4	79.0	88.9
DotSCN(c) SUB.	**62.5**	**83.0**	**89.8**	**95.5**	**70.4**	**79.0**	**84.0**	**90.1**

TABLE IV

PERFORMANCE EVALUATION ON DIFFERENT FUSION METHODS ON **DUKEMTMC GROUP** AND **ROAD GROUP**. THE SUFFIX 's' REPRESENTS THE RESULTS WITH SINGLE REPRESENTATIONS, AND THE SUFFIX 'c' REPRESENTS THE RESULTS WITH COUPLE REPRESENTATIONS, BOTH EXTRACTED BY RESNET50 BACKBONE, 'EQ.' DENOTES THE FUSION RESULTS OBTAINED WITH EQUAL WEIGHTS, AND 'AD.' DENOTES THE FUSION RESULTS OBTAINED WITH THE ADAPTIVE WEIGHT IN (3)

Method	DukeMTMC Group				Road Group			
	CMC-1	CMC-5	CMC-10	CMC-20	CMC-1	CMC-5	CMC-10	CMC-20
DotSCN(s)	80.7	89.8	94.3	96.6	82.7	93.8	**96.3**	**98.8**
DotSCN(c)	62.5	83.0	89.8	95.5	70.4	79.0	84.0	90.1
DotSCN EQ.	80.7	89.8	94.3	96.6	82.7	90.1	92.6	95.1
DotSCN AD.	**86.4**	**98.8**	**98.8**	**98.8**	**84.0**	**95.1**	**96.3**	**98.8**

Fig. 7. **Two visual examples showing the effectiveness of couple representations.** These two examples are selected from the results of **DukeMTMC Group** and **Road Group**, respectively. The first column shows the probe image. The following columns show the top three matches. The samples highlighted by red bounding boxes indicate the negative results, while those highlighted by green bounding boxes indicate the positive results. 'Single' and 'Couple' represent that the results are obtained respectively by the single or couple representation learning. These two examples show that, in some cases that the single representations lose their effectiveness, the couple representations take their responsibility and offer better predictions.

operation 'DotSCN(c) SUB.' and (2) addition operation 'DotSCN(c) ADD.'. Table III shows that the subtraction-based couple representation learning 'DotSCN(c) SUB' outperforms the addition-based learning 'DotSCN(c) ADD'.

D. Effectiveness of Adaptive Feature Fusion

In our method, the final results are generated by fusing results of single and couple representations with (3).

The fusion weights indicate the importance of the representations for the given probe image. To evaluate the proposed adaptive feature fusion 'DotSCN AD.' of single and couple representations, we compare it with (1) equal-weight feature fusion 'DotSCN EQ.', (2) transferred single representations only 'DotSCN(s)', and (3) transferred couple representations only 'DotSCN(c)'. Table IV shows that the adaptive fusion scheme presented in Section III-C outperforms the other compared methods in all CMC metrics. In contrast, the equal-weight fusion performs the same as transferred single representations on **DukeMTMC Group**, but slightly worse on **Road Group**. The results demonstrate the effectiveness of the adaptive weighting scheme guided by the enclosed areas of the rank-distance curves associated with the single and couple features.

Specifically, to investigate how our method assigns a higher weight to a useful representation, we visualize some typical examples in Fig. 8. The results of $id = 14$ show an example for which both the single and couple representations work well, since one person in the group wears a yellow coat, which is salient making the single and couple representations both discriminative. The results of $id = 41$ show an example for which the single representation does not work well, since all the persons in the group wear clothes in very common colors. By contrast, the results of $id = 10$ show an example for which the couple representation fails, as the group members wear very similar clothes, making the discrepancy of different persons small. These examples demonstrate that with the proposed online feature fusion scheme, a larger weight will be assigned to the more salient/discriminative representation.

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

Table V compares the performances of our method and several state-of-the-art methods on **DukeMTMC Group** and **Road Group**. The compared methods include

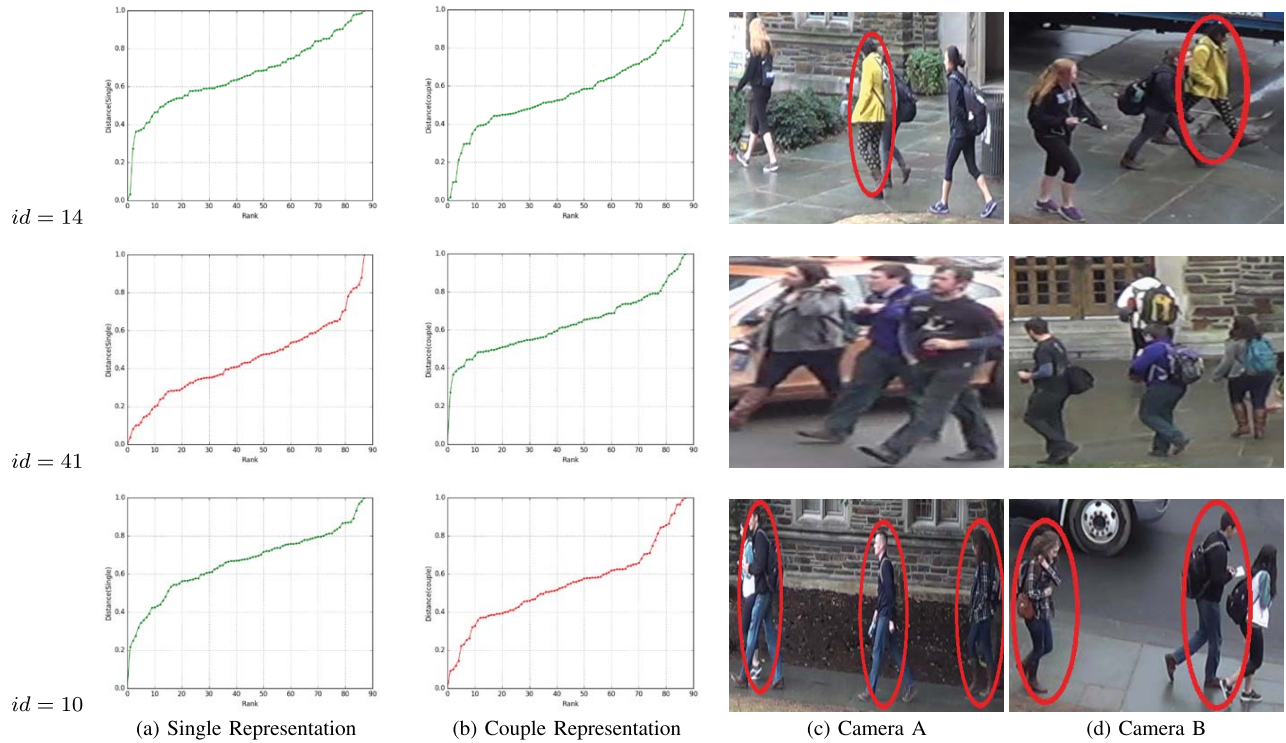|			
(a) Single Representation	(b) Couple Representation	(c) Camera A	(d) Camera B

Fig. 8. Examples showing the selection of single or couple representations. In each row, from left to right, four sub-figures are (a) the logarithmic rank-distance curve generated by the single representations of the given probe image, (b) the logarithmic rank-distance curve generated by the couple representations of the given probe images, (c) the given probe image from camera A, and (d) the matched image of the same group from camera B, respectively. Note that the red curve in (a) indicates that the images (c) and (d) are not perfectly matched by using single representation, whereas the red curve in (b) indicates that the images (c) and (d) are not perfectly matched by using couple representation.

TABLE V

COMPARISON WITH THE STATE-OF-THE-ART G-REID METHODS ON **DUKEMTMC GROUP** AND **ROAD GROUP**

Method	DukeMTMC Group				Road Group			
	CMC-1	CMC-5	CMC-10	CMC-20	CMC-1	CMC-5	CMC-10	CMC-20
CRRRO-BRO [24]	9.9	26.1	40.2	64.9	17.8	34.6	48.1	62.2
Covariance [23]	21.3	43.6	60.4	78.2	38.0	61.0	73.1	82.5
PREF [10]	22.3	44.3	58.5	74.4	43.0	68.7	77.9	85.2
BSC+CM [25]	23.1	44.3	56.4	70.4	58.6	80.6	87.4	92.1
MGR [11]	47.4	68.1	77.3	87.4	72.3	90.6	94.1	97.5
DotGNN [26]	53.4	72.7	80.7	88.6	74.1	90.1	92.6	98.8
DotSCN (Ours)	**86.4**	**98.8**	**98.8**	**98.8**	**84.0**	**95.1**	**96.3**	**98.8**

CRRO-BRO [24], Covariance [23], PREF [10], BSC+CM [25] and MGR [11]. The results show that our method significantly outperforms all existing G-ReID methods, thanks to the proposed transferred representation learning and adaptive feature fusion. In particular, compared with the best state-of-the-art method, our method achieves improvements by 11.7% CMC-1 on **Road Group** and 39.0% CMC-1 on **DukeMTMC Group**.

To evaluate the effectiveness of the proposed group representation that adaptively fuses single and couple features, we compare it with the graphical group representation model proposed in our previous work [26] that employs a graph neural network (GNN) to represent a group of individuals. The GNN-based framework in [26] involves a graph generator for constructing the pool of graph samples with the individuals' representations as nodes and a GNN model trained on the pool of graph samples for classifying the group IDs. The GNN plays the role of fusing all individuals; features together.

We implement the 'DotGNN' model that employs the GNN-based group representation model in [26] along with the domain-transfer method used in this work. As demonstrated in Table V, the adaptive single and couple feature fusion of DotSCN achieves significantly better performance than the GNN-based group representation of DotGNN in all CMC metrics. Specifically, DotSCN achieves CMC-1 performance improvements by 33.0% and 9.9%, respectively, on **DukeMTMC Group** and **Road Group** compared to Dot-GNN, because additional multi-person relations (*e.g.*, triplets

and quadruplets) other than the couple representation may contain too many similar features, making the GNN difficult to distinguish. On the contrary, the couple representation can capture more unique features, effectively complementing the single representation to improve the G-ReID performance.

V. CONCLUSION

In this article, we addressed an important but rarely studied problem: group re-identification. We have proposed a domain-transferred representation learning and a couple representation learning scheme to respectively overcome the two major challenges with group re-identification: the limited training data challenge and the membership and layout change challenge. We have also proposed an adaptive fusion method to combine the single and couple representations to achieve better group re-identification. Our experimental results have confirmed a performance leap from the relevant state-of-the-art techniques in the area.

REFERENCES

[1] S. Li, M. Shao, and Y. Fu, "Cross-view projective dictionary learning for person re-identification," in *Proc. Int. Joint Conf. Artif. Intell.*, Jun. 2015, pp. 2155–2161.

[2] Z. Wang *et al.*, "Zero-shot person re-identification via cross-view consistency," *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 260–272, Feb. 2016.

[3] Z. Wang, R. Hu, Y. Yu, J. Jiang, C. Liang, and J. Wang, "Scale-adaptive low-resolution person re-identification via learning a discriminating surface," in *Proc. Int. Joint Conf. Artif. Intell.*, Jul. 2016, pp. 2669–2675.

[4] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 2194–2200.

[5] X. Zhong, M. Feng, W. Huang, Z. Wang, and S. Satoh, "Poses guide spatiotemporal model for vehicle re-identification," in *Proc. Int. Conf. Multimedia Modeling*, 2019, pp. 426–439.

[6] L. Zhang, F. Liu, and D. Zhang, "Adversarial view confusion feature learning for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Jun. 16, 2020, doi: 10.1109/TCSVT.2020.3002956.

[7] X. Liu, S. Bi, S. Fang, and A. Bouridane, "Bayesian inferred self-attentive aggregation for multi-shot person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3446–3458, Oct. 2020.

[8] Y. Huang, J. Xu, Q. Wu, Y. Zhong, P. Zhang, and Z. Zhang, "Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3459–3471, Oct. 2020.

[9] R. Zhu, J. Fang, H. Xu, H. Yu, and J. Xue, "DCDLearn: Multi-order deep cross-distance learning for vehicle re-identification," 2020, *arXiv:2003.11315*. [Online]. Available: http://arxiv.org/abs/2003.11315

[10] G. Lisanti, N. Martinel, A. Del Bimbo, and G. L. Foresti, "Group re-identification via unsupervised transfer of sparse features encoding," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2449–2458.

[11] W. Lin *et al.*, "Group re-identification with multi-grained matching and integration," *IEEE Trans. Cybern.*, early access, Jun. 11, 2019, doi: 10.1109/TCYB.2019.2917713.

[12] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero-and homogeneously," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 172–188.

[13] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5157–5166.

[14] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian, "Query-adaptive late fusion for image search and person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1741–1750.

[15] Z. Huang, Z. Wang, T.-Y. Hung, S. Satoh, and C.-W. Lin, "Group re-identification via transferred representation and adaptive fusion," in *Proc. IEEE 5th Int. Conf. Multimedia Big Data (BigMM)*, Sep. 2019, pp. 128–132.

[16] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159.

[17] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1249–1258.

[18] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3960–3969.

[19] 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.

[20] F. Zhu, X. Kong, L. Zheng, H. Fu, and Q. Tian, "Part-based deep hashing for large-scale person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4806–4817, Oct. 2017.

[21] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2860–2871, Jun. 2019.

[22] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose-invariant embedding for deep person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4500–4509, Sep. 2019.

[23] Y. Cai, V. Takala, and M. Pietikainen, "Matching groups of people by covariance descriptor," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2744–2747.

[24] W.-S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 1–23.

[25] F. Zhu, Q. Chu, and N. Yu, "Consistent matching based on boosted salience channels for group re-identification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 4279–4283.

[26] Z. Huang, Z. Wang, W. Hu, C.-W. Lin, and S. Satoh, "DoT-GNN: Domain-transferred graph neural network for group re-identification," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1888–1896.

[27] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

[28] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.

[29] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3722–3731.

[30] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," 2016, *arXiv:1611.02200*. [Online]. Available: http://arxiv.org/abs/1611.02200

[31] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 994–1003.

[32] Q. Wang, M. Chen, F. Nie, and X. Li, "Detecting coherent groups in crowd scenes by multiview clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 46–58, Jan. 2020.

[33] C.-Y. Weng, W.-T. Chu, and J.-L. Wu, "RoleNet: Movie analysis from the perspective of social networks," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 256–271, Feb. 2009.

[34] C.-M. Tsai, L.-W. Kang, C.-W. Lin, and W. Lin, "Scene-based movie summarization via role-community networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 11, pp. 1927–1940, Nov. 2013.

[35] 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. Eur. Conf. Comput. Vis. Workshop*, 2016, pp. 17–35.

[36] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.

[37] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

Ziling Huang (Student Member, IEEE) received the B.S. degree in electrical engineering from North China Electric Power University, Beijing, China, in 2015, and the master's degree in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan, in 2020. She is currently pursuing the Ph.D. degree with the Department of Information and Communication Engineering, Graduate School of Information Science and Technology, The University of Tokyo. She was an Intern Student at the National Institute of Informatics, Tokyo, Japan, in 2019, and ByteDance, Beijing, from 2019 to 2020. Her research interests include computer vision and machine learning.
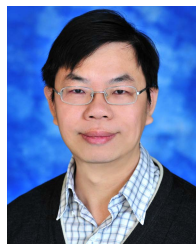
Zheng Wang (Member, IEEE) received the B.S. and M.S. degrees from Wuhan University, Wuhan, China, in 2006 and 2008, respectively, and the Ph.D. degree from the National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, in 2017. He was a JSPS Fellowship Researcher with the Shin'ichi Satoh's Lab, National Institute of Informatics, Japan, from 2018 to 2020. He is currently a Project Assistant Professor with the Research Institute for an Inclusive Society Through Engineering (RIISE) and the Graduate School of Information Science and Technology at The University of Tokyo. His research interests include person re-identification and instance search. He received the Best Paper Award at the 15th Pacific-Rim Conference on Multimedia (PCM 2014) and the 2017 ACM Wuhan Doctoral Dissertation Award.

Chung-Chi (Charles) Tsai (Member, IEEE) received the B.S. degree from National Tsing Hua University, Hsinchu, Taiwan, in 2009, the M.S. degree from the University of California at Santa Barbara, Santa Barbara, CA, USA, in 2012, and the Ph.D. degree from Texas A&M University at College Station, College Station, TX, USA, in 2018, all in electrical engineering. He attended a one-year exchange program at The University of New Mexico, Albuquerque, NM, USA, in 2007, and also participated in the summer internship with MediaTek in the summer of 2013/2015/2016. He is currently a Senior System Engineer for Camera ISP with Qualcomm Technologies, Inc., San Diego, CA, USA. His research interests include image processing, computational photography, and computer vision.

Shin'ichi Satoh (Member, IEEE) received the B.E. degree in electronics engineering and the M.E. and Ph.D. degrees in information engineering from The University of Tokyo, Tokyo, Japan, in 1987, 1989, and 1992, respectively. He has been a Full Professor with the National Institute of Informatics, Tokyo, since 2004. He was a Visiting Scientist with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA, from 1995 to 1997. His current research interests include image processing, video content analysis, and multimedia databases.

Chia-Wen Lin (Fellow, IEEE) received the Ph.D. degree in electrical engineering from National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2000.

He was with the Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan, from 2000 to 2007. Prior to joining academia, he worked for the Information and Communications Research Laboratories, Industrial Technology Research Institute, Hsinchu, from 1992 to 2000. He is currently a Professor with the Department of Electrical Engineering and the Institute of Communications Engineering at NTHU. He is also the Deputy Director of the AI Research Center, NTHU. His research interests include image and video processing, computer vision, and video networking.

Dr. Lin has served as a Distinguished Lecturer of the IEEE Circuits and Systems Society from 2018 to 2019, a Steering Committee Member of the IEEE TRANSACTIONS ON MULTIMEDIA from 2014 to 2015, and the Chair of the Multimedia Systems and Applications Technical Committee of the IEEE Circuits and Systems Society from 2013 to 2015. His articles received the Best Paper Award of IEEE VCIP 2015, Top 10% Paper Awards of IEEE MMSP 2013, and the Young Investigator Award of VCIP 2005. He received the Outstanding Electrical Professor Award presented by the Chinese Institute of Electrical Engineering in 2019 and the Young Investigator Award presented by the Ministry of Science and Technology, Taiwan, in 2006. He is the Chair of the Steering Committee of IEEE ICME. He has been serving as the President of the Chinese Image Processing and Pattern Recognition Association, Taiwan, since 2019. He has served as the Technical Program Co-Chair of IEEE ICME 2010 and IEEE ICIP 2019 and the General Co-Chair of IEEE VCIP 2018. He has served as an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE MULTIMEDIA, and *Journal of Visual Communication and Image Representation*.