SOCIAL INTERACTION MODELING FOR GROUP RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Group Re-identification (G-ReID) focus on associating group images that contain the same members across different camera views. The key challenge is that identity differentiation and position differentiation in group topology structure changes are difficult to capture. According to the social psychology principles, we found that the core members are more likely to remain in the group with smaller position changes, and peripheral members are more likely to have significant position changes or even fade out of the group. To this end, we propose a novel social interaction modeling (SIM), which treats group as a social interaction field, explore more authentic and robustness group features through the member differentiation. The member differentiation contains identity and position differentiation. Our method constructs the social interaction calculation module (SICM) to capture the member differentiation in fields, and implements identity differentiation and position differentiation by the social prior attention mechanism (SPAM) and social layout variation module (SLVM), respectively. A large number of experiments have been conducted on three available datasets show that the proposed method SIM is effective, and outperforms all previous state-of-the-art methods, surpassing the baseline on Rank1/mAP by up to 8.6%/9.6% on DukeGroup, 3.7%/2.7% on RoadGroup and 2.5%/2.9% on CSG. Code will be available on github.

1 INTRODUCTION

Group re-identification (G-ReID) aims to correctly associate group images containing the same members captured by different cameras with non-overlapping views. It is increasingly important in the security field. G-ReID typically deals with groups of 2-6 people, where images belonging to the same group category should contain at least 60% of the same members. G-ReID is more crucial and challenging than individual re-identification because people naturally exhibit group and social attributes, indicating that people prefer to move in groups and always engage in social interaction. Therefore, G-ReID needs to deel with member differentiation in group topology structure changes, which contains identity differentiation and position differentiation. Specifically, the identity differentiation means the importance(the chance of appearing in other images of same group id) of intra-group members varies due to different interaction probabilities in social interaction field, and position differentiation means the extent of position changes vary among intra-group members.

Although previous works (Zhang et al., 2024a; Yan et al., 2020; Zhang et al., 2022) based on deep learning to address the challenge of group topological structure changes, the performances are not satisfactory. The shortcomings are mainly due to the following two reasons: 1) Existing methods address the challenge from the perspective of the entire group distribution. As shown in Fig 1, the extracted group features are the undifferentiated features,which are shown as the pure pink triangles, lead to large intra-class distance. 2) The previous attention mechanism conducts undifferentiated learning for all group member features, lacking of specific focus. And previous layout modeling employs undifferentiated random affine transformations for each intra-group member, leading to many ineffective layout.

According to the social psychology principles (Latane et al., 1980; Lewin, 1943), complex social interactions can be represented as physical fields, and (Zhou et al., 2019) translated social features into a quantitative formula for interaction probability. Through focus these principles, we found that the core members are more likely to remain in the group with smaller position changes across dif-

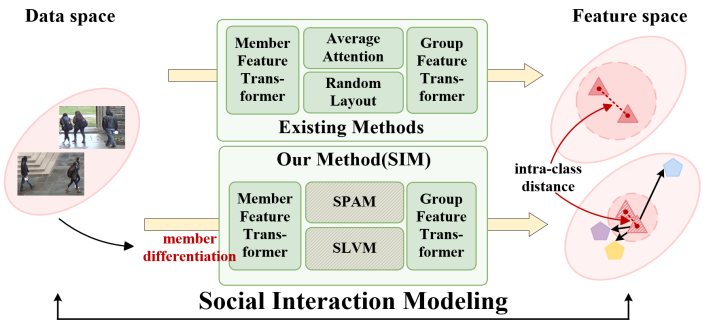


Figure 1: Existing methods versus social interaction modeling. Pure triangles and texture triangles represent the group feature mined without and with member differentiation. The pentagon represents the different member features that make up the group features. The dotted circle lines represent the boundaries within the class. The features extracted by our SIM have smaller intra-class distance due to consider the member differentiation.

Figure 2: member differentiation. Two images belong to the same group, and the numbers represent the average interaction probability. B-boxes of same color represent the same member.

ferent group images, and peripheral members are more likely to have significant positional changes or even fade out of the group images. The core members refer to the members with higher interaction probabilities relatively (e.g. members that be surrounded), while peripheral members refer to the members with lower interaction probabilities (e.g., unrelated pedestrians). As shown in Fig 2, members in red and orange b-box are with higher interaction probability, remain their core position, and have more chance appear in other same-group images, while members in green and blue b-box is with lower probability, one left the group, another has significant position changes.

In this paper, we propose a novel social interaction modeling method, which is motivated by the fact that based on the sociological fact that intra-group members have differentiation, which contains identity differentiation and position differentiation. The member differentiation cannot be erased, no matter how perfect the group entire distribution is. Therefore, member differentiation is a social attribute of the group that cannot disappear by perfecting the entire distribution of the group. The proposed social interaction modeling treats each group image as a social interaction field with member differentiation rather than an entire distribution, constructs the social interaction calculation module (SICM) to capture the member differentiation in fields, and digs out more authentic group features through the identity differentiation and position differentiation in member differentiation. Two modoules, the social prior attention mechanism (SPAM) and social layout variation module (SLVM), are designed to accomplish identity differentiation and position differentiation. As shown in Fig.1, the group features learned with social interaction modeling has samller intra-class distance, and consistent with the real-world distribution. Modeling and training this differentiation can obtain smaller intra-class distance and more authentic and robust group feature representations.

Specifically, the proposed SICM defines a normalized variable $\hat{p}$ to reflect the member differentiation. A standard form $\hat{p}$ is constructed with the following properties. First, each group image has a specific $\hat{p}$ due to the different social interaction features among intra-group members. Second, the accuracy of the $\hat{p}$-value is mainly affected by distance. SICM extracts social interaction features of intra-group members in the group images. The social interaction features contains distance d, orientation θ, openness o, which are extracted by using data annotations and group images.

SPAM foucs on accomplish identity differentiation. Because the importance (the chance of appearing in other group images) of intra-group members varies from each other, SPAM adjusts the weight of attention to different members during group feature learning. To this end, a new attention weight allocation mechanism is designed to achieve core member mining and enhance group feature learning, achieve identity-differentiated learning of features.

SLVM aims to adress position differentiation. Due to the extent of position variation varies from intra-group members, SLVM models more realistic dynamic layout variations. Thus, a learnable position variation matrix $\triangle D$ is employed. For a group with j members, the j-th column of $\triangle D$

is the differentiation position variation of the j-th member. While retaining a certain degree of freedom through weighted fusion with random affine vectors. SLVM accomplish a new layout modeling strategy to conduct more realistic layout modeling and explore potential layout changes, achieve position-differentiated learning of features.

Our main contributions are summarized as follows:

- We firstly introduce the social psychology principles into G-ReID task, and accomplished member differentiation.
- We propose the social interaction modeling (SIM) method, which treats each group image as a social interaction field instead of an entire distribution, and constructs the social interaction calculation module (SICM) to capture the member differentiation in fields. Social interaction modeling aimes to explore more authentic group features through identity differentiation and position differentiation, which is achieved by the proposed social prior attention mechanism (SPAM) and social layout variation module (SLVM).
- Our SIM achieves 96.1%/95.5%, 92.6%/94.4%, and 83%/89% Rank-1/mAP on CSG, RoadGroup and DukeGroup datasets, outperforming all of the state-of-the-art method.

2 Related Work

Person Re-identification. Person re-identification (ReID) aims to associate individual pedestrians across non-overlapping views in camera networks. In recent years, numerous deep learning-based methods (Meng et al., 2019; 2021; Yan et al., 2020; Rao & Miao, 2023; Wang et al., 2024; Zhang et al., 2024b; Peng et al., 2023; Guo et al., 2024) have made significant progress in this field, including extracting more discriminative features and designing more suitable metrics. For instance, FSAM (Hong et al., 2021) proposed a dual-stream framework to extract fine-grained body features, while AGW (Ye et al., 2021) introduced a weighted regularized triplet metric learning method. However, person ReID methods primarily focus on individual pedestrians, overlooking the more intricate group-level interactions and layout dynamics that are pivotal for GReID.

Group Re-identification. Compared to ReID, research on G-ReID remains relatively scarce, with only a few pioneering works attempting to address this task. Early approaches (Zheng et al., 2011; Cai et al., 2010; Ristani et al., 2016; Lisanti et al., 2017) treated entire images as model inputs and directly extracted group features. Since these methods relied on handcrafted features and considered background information, their performance was unsatisfactory. Subsequently, CNN-based works (Mei et al., 2019; 2020; 2021) became mainstream, where group members were first cropped before extracting group features. For example, DotGNN (Huang et al., 2019) transferred styles from single-person ReID datasets to group ReID datasets and employed graph neural networks to learn group graph features. GCGNN (Zhu et al., 2020) constructed a graph neural network framework based on spatial K-nearest neighbor graphs, achieving neighborhood aggregation through node in-degree and spatial relationship attributes. MACG (Yan et al., 2020) proposed a multi-level attention contextual graph model to leverage visual context information among group members. In recent years, Vision Transformer-based architectures have gained popularity. UMSOT (Zhang et al., 2022; 2024a) introduced a second-order Transformer architecture to construct group features, incorporating uncertainty modeling of group member number and position. PBSOT (Zhang et al., 2025) prosed a parallel branches-based transformer with layout-guided occlusion mitigation, enhances robustness by strengthening the sampling of overlapping parts and fusing global features with local features. But the existing methods were from the perspective of the entire group distribution, not notice the members differentiation of intra-group members.

Social Interaction. (Bolotta & Dumas, 2022) identified social interaction as a key area for future AI research in 2022, revealing that certain visual primitive features of social behavior discovered by cognitive psychologists enhance computer vision systems' ability to recognize interactions. VAGS (Leach et al., 2014) proposed a visual attention-guided social group detection framework that improved motion-based social group estimation by inferring pedestrian gaze directions. SIFM (Zhou et al., 2019) uses the avatars generated by VR, a mathematical model is established to calculate the interaction probability, and the results are psychologically explained. However, these work did not apply social psychology principles into GReID task, Our SIM method made an attempt.

3 METHOD

In this section, we firstly introduce the SICM, and then we describe the proposed SPAM and SLVM of social interaction modeling. Fig 3 illustrates the method in detail. The purpose of GReID is to match groups composed of the same members across different camera views. For the k-th group, x_k and y_k^g are the group image and id, respectively, and b_k and y_k^p are the bounding box annotation and member id for each member in x_k, respectively.

3.1 SOCIAL INTERACTION CALCULATION MODULE (SICM)

In this paper, SICM aims to capture the member differentiation with normalized variable $\hat{p_i}$ in social interaction fields. The key issue is extract social interaction features, and calculate interaction probabilities of each group image. For the k-th group image, we treats it as a social interaction field $S_k = \left\{S_{ij}^k\right\}$, member i and j have their subfield S_{ij}^k. Binary variable $z_{ij} \in \{0, 1\}$,following a Bernoulli distribution, determines whether pedestrians i and j in same social interaction field:

$$p\left(z_{ij} = 1 \mid S_{ij}^k\right) = \delta\left(S_{ij}^k\right). \tag{1}$$

To calculate the interaction probability p_{ij} between member i and j, we now extract social interaction features: distance, orientation and openness. Specifically, we utilize b-boxes b_{ki}, b_{kj} to calculate distance d_{ij}. The distance between the i-th and j-th members based on their bboxes:

$$d_{ij} = \frac{1}{\gamma}\left\|\psi(b_{ki}^{mid}, b_{kj}^{mid})\right\|, \tag{2}$$

where b_{ki}^{mid} and b_{kj}^{mid} enote the bottom midpoints of b-boxes, and γ is a scaling factor. $\psi(x_1, x_2) = \phi(x_1) - \phi(x_2)$ is constructed to compensate for field-of-view discrepancies, where ϕ denotes a perspective transformation function (Zhang, 2021).

Then, we utilize multiple frameworks merging such as Mediapipe (Lugaresi et al., 2019), AlphaPose (Fang et al., 2022), HigherHRNet (Cheng et al., 2020) to extract skeletal keypoints $\boldsymbol{q}_i$ from image of cropped member x_{ki} to compute relative orientations θ_{ij} and define pose-openness o_i, o_j. We compute the relative angle:

$$\theta_{ij} = arccos\frac{\boldsymbol{v}_i^{\perp} \cdot \boldsymbol{d}_{ij}}{\left\|\boldsymbol{v}_i^{\perp}\right\| \cdot \left\|\boldsymbol{d}_{ij}\right\|}, \tag{3}$$

where $\boldsymbol{q}_i^{ls}$ and $\boldsymbol{q}_i^{rs}$ are the left and right shoulder keypoints of the i-th pedestrian, respectively. $\boldsymbol{v}_i = \boldsymbol{q}_i^{rs} - \boldsymbol{q}_i^{ls}$ is shoulder vector of the i-th member, $\boldsymbol{v}_i^{\perp}$ represents the orientation vector.

The pose-openness degree o_i is defined as:

$$o_i = \begin{cases} 1, & \text{if } \zeta_{ub} > \zeta_1, \\ -1, & \text{if } \boldsymbol{q}_i^{ws} \times \boldsymbol{q}_i^{bd} > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

where $\zeta_{ub} = \left\langle \boldsymbol{q}_i^{up}, \boldsymbol{q}_i^{bd} \right\rangle$ is the angle between the upper arm $\boldsymbol{q}_i^{up}$ and body $\boldsymbol{q}_i^{bd}$, ζ_1 is a threshold defined 45 degree. $\boldsymbol{q}_i^{ws}$ is forearm, and $\times$ is the cross product denotes vector outer product. Indicates that o_i equal to 1 when upper arm is spread out, and o_i equal to -1 when forearm tightens inward towards the body. Now we have social interaction features.

Then we calculate the p_{ij}. The subfield intensity S_{ij}^k and interaction probability satisfy the formula:

$$\begin{cases} S_{ij}^k = f(d_{ij}, \theta_{ij}, o_i, o_j) \cdot \delta\left(S_{ij}^k\right) \\ p_{ij} = 1 - exp(-S_{ij}^k/\lambda)^b \end{cases} \tag{5}$$

where function f is a symmetric function for i and j, parameters λ, b are givern in (Zhou et al., 2019). $P = \{p_{ij}\}_{i=1}^N$ is a symmetric matrix, because $S_{ij}^k = S_{ji}^k$. The average interaction probability $\bar{p_i}$, and normalized average interaction probability $\hat{p_i}$ can be describe as:

$$\begin{cases} \bar{p_i} = \frac{1}{N-1}\sum_{j=1, j\neq i}^{N} p_{ij} \\ \hat{p_i} = \bar{p_i} / \sum_{i=1}^{N} \bar{p_i} = \frac{1}{N-1}\sum_{j=1, j\neq i}^{N} p_{ij} / \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N} p_{ij} \end{cases} \tag{6}$$

where N is the group size. For the i-th group $\hat{\boldsymbol{p}_k^g} = \{\hat{p_i}\}_{i=1}^{N_k} \in \mathbb{R}^{N_k}$.

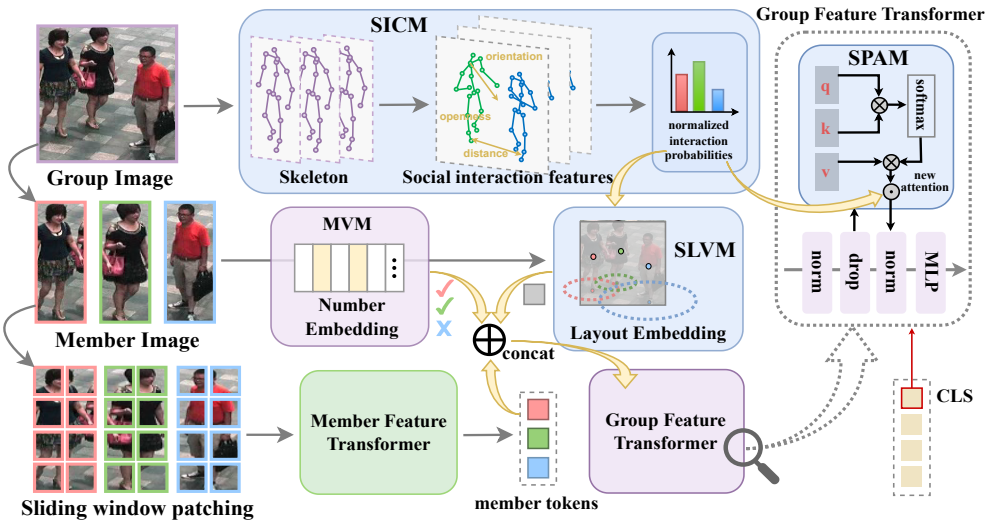


Figure 3: The pipline of the proposed SIM. The blue squares stands for the module proposed by us, purple and green squares represent the existing group and member modules, MVM is from our baseline. The grey arrows indicate data flow, and the yellow arrows indicate cross-module usage. The small gray squares represent layout.

3.2 Social Prior Attention Mechanism (SPAM)

SPAM aims accomplish identity differentiation. To this end, a new attention weight allocation mechanism is designed to achieve core member mining and enhance group feature learning. Specifically, higher attention weights assigns to core members to enhance the learning of core members' tokens.

The i-th member's token t_i^p of k-th group are extracted by ViT, $t_i^p = ViT(x_{ki})$, group feature is concatenated from members token, $t_k^g = [t_1^p, t_2^p, ..., t_{N_k}^p]$. The input of group vision transformer (GViT) are $X = [t_m^g, t_{m+1}^g, ..., t_{m+B-1}^g] \in \mathbb{R}^{B\times N\times C}$, where B denotes the batch size, N is the number of group members, C represents the feature dimension, p and g represents person and group, m due to current batch. The query (Q), key (K), value (V) are obtained through linear transformations: $Q = XW_Q, K = XW_K, V = XW_V$, where $W_Q, W_K, W_V \in \mathbb{R}^{C\times d}$ are learnable parameters, and d is the attention head dimension. The original attention weights are calculated as:

$$A_{raw} = (Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V. \tag{7}$$

The new attention is described as:

$$A_{social} = A_{raw} \odot \hat{\boldsymbol{P}}^g, \tag{8}$$

where $\hat{\boldsymbol{P}}^g = [\hat{\boldsymbol{P}}_1^g, \hat{\boldsymbol{P}}_2^g, ..., \hat{\boldsymbol{P}}_d^g] \in \mathbb{R}^{B\times C\times d}$, and $\hat{\boldsymbol{P}}_d^g = [\hat{\boldsymbol{p}}_m^g, \hat{\boldsymbol{p}}_{m+1}^g, ..., \hat{\boldsymbol{p}}_{m+B_d-1}^g] \in \mathbb{R}^{B\times C}$, and $\odot$ represents Hadamard product. The updated features Z are generated through linear projection: $Z = A_{social}VW_O$, where $W_O \in \mathbb{R}^{d\times C}$ is a learnable projection matrix. By integrating $\hat{\boldsymbol{P}}^g$ and dimension alignment, SPAM optimizes attention weights, accomplish identity differentiation.

3.3 Social Layout Variation Module (SLVM)

SLVM aims to dress position differentiation. This module accomplish a new layout modeling strategy to conduct more realistic layout modeling and explore potential layout changes D_{final}. It constructs a learnable position variation matrix $\triangle D$, which: restricts layout variation ranges for core members and expands layout variation ranges for peripheral members. This mechanism simulates realistic sociological layout variations that incorporate position differentiation.

Original layout coordinates of group image $D_{ori} = [(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)] \in \mathbb{R}^{N\times 2}$ represents the center coordinates of each member in the image, (x_i, y_i) represents the position of i-th member.

The existing methods treat the group layout as a entirety and apply random affine transformation, which without considering the positional discrimination. The transformed layout after random affine transformation is given by: $D_{random} = RD_{ori}+b$, where $R \in \mathbb{R}^{2\times 2}$ is random affine matrix, $b \in \mathbb{R}^2$ is translation vector.

For the position variation of i-th member of the group, our layout modeling can be described as:

$$\triangle d_i = \alpha \left(\sum_{j\neq i} \hat{p_j}(d_j - d_i) \right) + (1-\alpha) r_i, \tag{9}$$

where d_i and d_j are the central position of i-th and j-th member's bbox, and $\sum_{i\neq j} \hat{p_i}(d_i - d_j)$ represents position differentiation. Hyperparameter $\alpha \in [0, 1]$ is a balancing coefficient that weights prior differentiation knowledge against data augmentation. r_j is a random perturbation vector, can maintain a certain degree of positional freedom for members.

The offset $\triangle d_j$ consists of two components: 1) Structure-aware offset driven by social weights to preserve spatial proximity for strongly connected pedestrians. 2) Random perturbation to introduce diversity.

Specifically, $(d_i - d_j)$ represents distance in real, $\hat{p_i}$ is normalized average interaction probability, which reflect member differentiation. Therefore, members with higher interaction probability with the i-th member can better maintain the distance between them and are accompanied by smaller positional variation. Meanwhile, members with lower interaction probability with the i-th member cannot maintain the distance between them and will generate greater layout variation.

The final layout is computed as:

$$D_{final} = D_{ori} + \triangle D, \tag{10}$$

where $\triangle D \in \mathbb{R}^{N\times 2}$ is the offset matrix for all intra-group members. The updated features Z are fused with the layout information D_{final} for Transformer encoder generates the group feature representation.

3.4 Loss function

Our feature is learning supervised by the person identity and triplet loss function.

$$\mathcal{L}_{ID} = -\frac{1}{P} \sum_{j=1}^{P} \sum_{i=1}^{C} y_{ji} log(\widehat{y}_{ji}), \tag{11}$$

where P represents the total member number of the current batch, C represents the total member classes, the indicator function $y_{ji} = 1(j = i)$ equals to 1 when the j-th member belongs to the i-th class, and $\widehat{y}_{ji}$ is the prediction of network about the j-th member belongs to the i-th class.

$$\mathcal{L}_{Tri} = \frac{1}{P} \sum_{i=1}^{P} max(d(f_i, f_i^+) - d(f_i, f_i^-) + m, 0), \tag{12}$$

where $d(\cdot, \cdot)$ represents the distance function between two features such as the Euclidean distance, $f_i/f_i^+/f_i^-$ represent the anchor/hard positive/hard negative feature in the current batch, and m is the hyper-parameter of margin. $\mathcal{L}_p$ is person loss $\mathcal{L}_p = \mathcal{L}_{ID} + \mathcal{L}_{Tri}$.

The loss function $\mathcal{L}_g$ of a second-order token (Zhang et al., 2024a) is also composed of the group identity and triplet loss, which is similar to the $\mathcal{L}_{ID}$ and $\mathcal{L}_{Tri}$.Overall, the whole loss function is described as:

$$\mathcal{L} = \mathcal{L}_p + \mathcal{L}_g. \tag{13}$$

Table 1: The proposed method is compared with state-of-the-art approaches on CSG, RoadGroup, and DukeGroup datasets. The comparative methods are divided into two categories: hand-crafted and deep learning-based methods. The best and second-best results are highlighted in bold and underlined, respectively. Reported metrics include Rank-1, Rank-5, Rank-10 and mAP (%).

Method	Publication	CSG				RoadGroup				DukeGroup			
		Rank1	Rank5	Rank10	mAP	Rank1	Rank5	Rank10	mAP	Rank1	Rank5	Rank10	mAP
CRRRO-BRO	BMVC 2009	10.4	25.8	37.5	-	17.8	34.6	48.1	-	9.9	26.1	40.2	-
Covariance	ICPR 2010	16.5	34.1	47.9	-	38.0	61.0	73.1	-	21.3	43.6	60.4	-
BSC-CM	ICIP 2016	24.6	38.5	55.1	-	58.6	80.6	87.4	-	23.1	44.3	56.4	-
PREF	ICCV 2017	19.2	36.4	51.8	-	43.0	68.7	77.9	-	30.6	55.3	67.0	-
LIMI	MM 2018	-	-	-	-	72.3	90.6	94.1	-	47.4	68.1	77.3	-
DotGNN	MM 2019	-	-	-	-	74.1	90.1	92.6	-	53.4	72.7	80.7	-
DotSNN	TCSVT 2019	-	-	-	-	84.0	95.1	96.3	-	-	-	-	-
GCGNN	TMM 2020	-	-	-	-	81.7	94.3	96.5	-	53.6	77.0	91.4	-
SVIGR	Neucom 2020	-	-	-	-	87.8	92.7	-	89.2	-	-	-	-
MGR	TCYB 2021	57.8	71.6	76.5	-	80.2	93.8	96.3	-	48.4	75.2	89.9	-
MACG	TPAMI 2023	63.2	75.4	79.7	-	84.5	95.0	96.9	-	57.4	79.0	90.3	-
SOT	AAAI 2022	91.7	96.5	97.6	90.7	86.4	96.3	98.8	91.3	72.7	88.6	93.2	78.9
UMSOT	IJCV 2024	93.6	97.3	98.3	92.6	88.9	95.1	**98.8**	91.7	74.4	89.4	93.9	79.4
PBSOT	ESWA 2025	94.5	97.1	97.9	93.9	91.3	**96.3**	98.7	93.3	82.7	92.6	95.1	88.1
Ours	-	**96.1**	**98.3**	**99.1**	**95.5**	**92.6**	**96.3**	97.5	**94.4**	**83.0**	**96.6**	**98.9**	**89.0**

4 EXPERIMENTS

4.1 DATASETS

The proposed SIM is evaluated on three G-ReID datasets: DukeGroup, RoadGroup (Lin et al., 2019)), and CSG (Yan et al., 2020). The DukeGroup dataset contains 354 images with 177 group classes. The RoadGroup dataset contains 324 images with 162 group classes. Following the protocol in (Yan et al., 2020), the training and test sets of DukeGroup and RoadGroup are randomly divided equally. The CSG dataset contains 3,839 images with 1,558 group classes, where 859/699 groups are allocated for training/testing. According to (Yan et al., 2020), test images are sequentially selected as probes, while all remaining images serve as the gallery. Additionally, CSG includes 5K extra group images in the gallery as distractors. For fair comparison, no additional datasets are used during training on any G-ReID dataset. Evaluation metrics include Rank-1, Rank-5, Rank-10 cumulative matching characteristics (CMC) and mean average precision (mAP).

4.2 DETAILS

Our baseline is UMSOT. The experiments are conducted on an 4080 GPU with Pytorch. Our model uses GViT (Zhang et al., 2022; 2024a) as backbone, pretrained on ImageNet (Deng et al., 2009). For input group images, we crop all member images using given bounding boxes and resize them to 256×128. During training, the random seed is fixed to 42, with random horizontal flipping (p=0.5) and random erasing applied. Each mini-batch samples 16 group identities, with 4 images selected per identity. We use SGD (Bottou, 2012) as optimizer. Training terminates after 400 epochs. A cosine annealing learning rate schedule is employed: initial rate 2e-3, minimum rate 1.6e-4. The learning rate for inter-member modules is multiplied by 0.1. Weight decay is set to 1e-4. Online hard mining is used for triplet loss (Hermans et al., 2017). During testing, no data augmentation or re-ranking is applied. Features are compared using Euclidean distance. Unless otherwise specified, all ablation studies, parameter analyses, and visualizations are conducted on the DukeGroup dataset.

4.3 PERFORMANCE

We compare SIM with existing methods on three available G-ReID datasets to demonstrate its superiority. We categorize existing methods into two groups:handcrafted G-ReID methods. and deep learning-based G-ReID methods, From the performance perspective, SIM is recognized as the state-of-the-art method among existing approaches. Two conclusions can be drawn from Table 1:

First, our SIM achieves strong performance on the CSG, RoadGroup, and DukeGroup datasets, surpassing baseline in both Rank-1 and mAP metrics. On the CSG dataset, SIM achieves 96.1%/95.5% (Rank-1/mAP), outperforming baseline by 2.5%/2.9 in Rank-1/mAP. On the RoadGroup dataset, SIM achieves 92.6%/94.4% (Rank-1/mAP), outperforming baseline by 3.7%/2.7% in Rank-1/mAP. On the DukeGroup dataset, SIM achieves 83.0%/89.0% (Rank-1/mAP), outperforming baseline by 8.6%/9.6% in Rank-1/mAP. These results demonstrate that SIM delivers performance gains across

Table 2: Ablation Experiments on the Social Interaction Module (SIM) (%).

SPAM	SLVM	CSG		RoadGroup		DukeGroup	
		Rank1	mAP	Rank1	mAP	Rank1	mAP
		93.57	92.62	88.89	91.73	74.42	79.40
✓		**96.11**	95.28	91.36	93.75	80.68	87.48
	✓	95.59	95.22	89.94	92.23	78.41	85.88
✓	✓	96.06	**95.49**	**92.59**	**94.35**	**82.95**	**89.02**

all datasets, confirms that the member differentiation of SIM is effectively, leading to significant improvements. Second, the performance of existing method remains unsatisfactory due to: Existing methods from the perspective of the entire group distribution, not considered the member differentiation in group topology structure changes. Unlike these methods, SIM construct a SICM to capture the member differentiation in social interaction fields. And the proposed SPAM achieve core member mining and enhance group feature learning, while SLVM accomplish a new layout modeling strategy to conduct more realistic layout modeling and explore potential layout changes, making SIM shorten the intra-class distance, enhanced the robustness of group features.

4.4 Ablation Study

Effect of SPAM and SLVM. The ablation experiments primarily demonstrate the impact of the SPAM and SLVM modules on social interaction modeling. We mainly analyze the results on the DukeGroup dataset, with similar conclusions observed on the other two datasets. As shown in Table 2, two conclusions can be drawn: First, each module individually improves performance when used separately. Compared to the baseline, SPAM increases Rank-1/mAP by +6.26%/+8.08%, while SLVM increases Rank-1/mAP by +3.99%/+6.48%. This indicates that these two modules respectively learn the identity and position differentiation, making SIM more discriminative. Second, when both modules are used together, the performance gains reach +8.53%/+9.62% in Rank-1/mAP, exceeding the sum of individual improvements. This demonstrates that social prior attention mechanism (SPAM) and the social layout variation module (SLVM) are two complementary aspects for mining group differentiated features. Using both SPAM and SLVM simultaneously enables the exploration of more authentic and robustness group features.

4.5 Parameter Analysis

Influence of α. The hyperparameter α controls SLVM learns potential group layouts by determining the ratio between the social interaction layout matrix and random affine matrix during training in SIM. When α is set too large, SLVM over-emphasizes layout variations of peripheral members, leading to overfitting and performance degradation. When α is set too small, SLVM fails to adequately explore potential layout features of peripheral members, resulting in insufficient model generalization and performance decline. As shown in Figure 4, we conducted separate hyperparameter experiments on CSG, RoadGroup and DukeGroup datasets to determine optimal values, since each dataset exhibits different social interaction fields distributions.

4.6 Visualization

Retrieval visualization. Figure 5 presents the top-3 retrieval visualizations comparing baseline UMSOT and our proposed SIM. The advantages of SIM are primarily demonstrated in two aspects: 1) Achieve core member mining and enhance group feature learning, achieve identity differentiated learning. UMSOT tends to retrieve gallery images with higher overall similarity to the query. When processing groups with less distinctive members (Rows 1, 2, and 4), SIM effectively focuses on core members of the group. 2) Conduct more realistic layout modeling and explore potential layout changes, achieve position differentiated learning. UMSOT does not emphasize layout variations. For groups in Rows 3 and 5, SIM better captures the topological changes of the group structure.

Feature visualization. Figure 6 illustrates the group feature visualization of our best eight classes in the training set when both UMSOT and our proposed SIM converge, and each group class contains only two images. Due to UMSOT's approach regards groups as entire distributions during the feature learning process, the intra-class distance is large. In contrast, our social interaction modeling

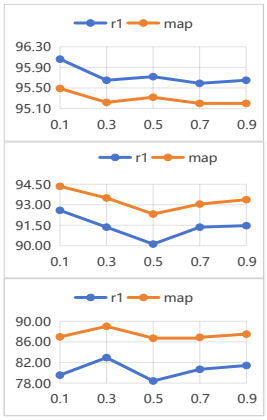


Figure 4: Parameter α on CSG, RoadGroup and DukeGroup.

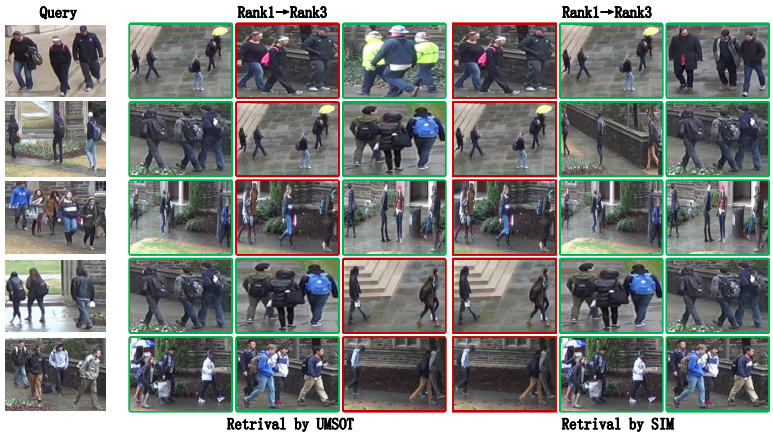
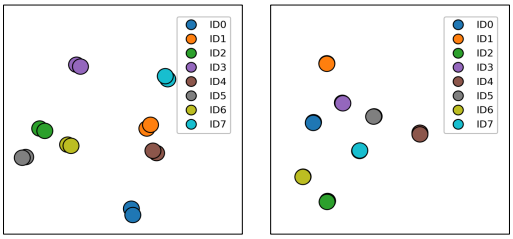


Figure 5: Top-3 Retrieval Visualization of UMSOT and SIM. Red/green bounding boxes indicate correct/incorrect matches, respectively. In the DukeGroup dataset, each query has only one correct match in the gallery.



(a) UMSOT (b) SIM

Figure 6: The feature visualization of the whole training set through t-SNE (Van der Maaten & Hinton, 2008). Each color represents a group class.

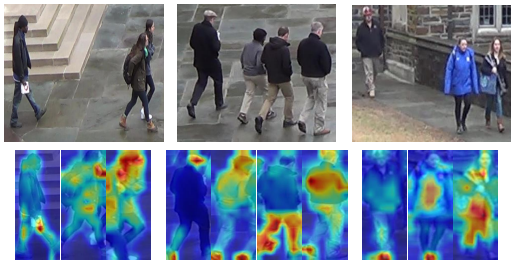


Figure 7: Heatmap visualization via Grad-CAM is applied to dataset images. Pedestrians with higher social interaction probability exhibit larger heatmap areas.

utilizes the member differentiation in learning group features, demonstrating: excellent intra-class consistency and strong inter-class differentiation.

Hotmap. We optimize group feature learning by adjusting the weights in the group feature Transformer attention mechanism–SPAM, and visualize individual member features combine group weight using Grad-CAM (Selvaraju et al., 2017) class activation heatmaps as shown in Figure 7. The results show that members with higher social interaction probability exhibit larger and more concentrated heatmap regions, indicating that the learning of group features is optimized and back-propagates to affect individual representations.

5 Conclusion

In this paper, we focus on the member differentiation in group topology structure changes in G-ReID, which contains identity and position differentiation. To solve this, we propose a novel social interaction modeling (SIM) method, which treats group as a social interaction field. First, our method constructs the social interaction calculation module (SICM) to capture the member differentiation in fields, and accomplish identity differentiation and position differentiation by the social prior attention mechanism (SPAM) and social layout variation module (SLVM), respectively, shorten the intra-class distance. Second, SPAM design a new attention weight allocation mechanism, to mine core member and enhance group feature learning. SLVM achieve a new layout modeling strategy to conduct more realistic layout modeling and explore potential layout changes. Finally, our proposed social interaction modeling (SIM) achieves state-of-the-art performance across multiple benchmark datasets, outperforming all existing methods. Limitations are in Appendix A due to the paper length.

REFERENCES

Samuele Bolotta and Guillaume Dumas. Social neuro ai: Social interaction as the "dark matter" of ai. *Frontiers in computer science*, 4:846440, 2022.

Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: tricks of the trade: second edition*, pp. 421–436. Springer, 2012.

Yinghao Cai, Valtteri Takala, and Matti Pietikainen. Matching groups of people by covariance descriptor. In *2010 20th International Conference on Pattern Recognition*, pp. 2744–2747. IEEE, 2010.

Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5386–5395, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009.

Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):7157–7173, 2022.

Wenxuan Guo, Zhiyu Pan, Yingping Liang, Ziheng Xi, Zhicheng Zhong, Jianjiang Feng, and Jie Zhou. Lidar-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17437–17447, 2024.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

Peixian Hong, Tao Wu, Ancong Wu, Xintong Han, and Wei-Shi Zheng. Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10513–10522, 2021.

Ziling Huang, Zheng Wang, Wei Hu, Chia-Wen Lin, and Shin'ichi Satoh. Dot-gnn: Domain-transferred graph neural network for group re-identification. In *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 1888–1896, 2019.

Bibb Latane, Steve Nida, et al. Social impact theory and group influence: A social engineering perspective. *Psychology of group influence*, pp. 3–34, 1980.

Michael JV Leach, Rolf Baxter, Neil M Robertson, and Ed P Sparks. Detecting social groups in crowded surveillance videos using visual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 461–467, 2014.

Kurt Lewin. Defining the'field at a given time.'. *Psychological review*, 50(3):292, 1943.

Weiyao Lin, Yuxi Li, Hao Xiao, John See, Junni Zou, Hongkai Xiong, Jingdong Wang, and Tao Mei. Group reidentification with multigrained matching and integration. *IEEE transactions on cybernetics*, 51(3):1478–1492, 2019.

Giuseppe Lisanti, Niki Martinel, Alberto Del Bimbo, and Gian Luca Foresti. Group re-identification via unsupervised transfer of sparse features encoding. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2449–2458, 2017.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

Ling Mei, Jianhuang Lai, Xiaohua Xie, Junyong Zhu, and Jun Chen. Illumination-invariance optical flow estimation using weighted regularization transform. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(2):495–508, 2019.

Ling Mei, Jianhuang Lai, Zhanxiang Feng, and Xiaohua Xie. From pedestrian to group retrieval via siamese network and correlation. *Neurocomputing*, 412:447–460, 2020.

Ling Mei, Jianhuang Lai, Zhanxiang Feng, and Xiaohua Xie. Open-world group retrieval with ambiguity removal: A benchmark. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 584–591. IEEE, 2021.

Jingke Meng, Sheng Wu, and Wei-Shi Zheng. Weakly supervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 760–769, 2019.

Jingke Meng, Wei-Shi Zheng, Jian-Huang Lai, and Liang Wang. Deep graph metric learning for weakly supervised person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6074–6093, 2021.

Jinjia Peng, Guangqi Jiang, and Huibing Wang. Adaptive memorization with group labels for unsupervised person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(10):5802–5813, 2023.

Haocong Rao and Chunyan Miao. Transg: Transformer-based skeleton graph prototype contrastive learning with structure-trajectory prompted reconstruction for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22118–22128, 2023.

Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pp. 17–35. Springer, 2016.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Runmin Wang, Zhenlin Zhu, Yanbin Zhu, Hua Chen, Yongzhong Liao, Ziyu Zhu, Yajun Ding, Changxin Gao, and Nong Sang. Dimgnet: A transformer-based network for pedestrian reidentification with multi-granularity information mutual gain. *IEEE Transactions on Multimedia*, 26: 6513–6528, 2024.

Yichao Yan, Jie Qin, Bingbing Ni, Jiaxin Chen, Li Liu, Fan Zhu, Wei-Shi Zheng, Xiaokang Yang, and Ling Shao. Learning multi-attention context graph for group-based re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):7001–7018, 2020.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021.

Quan Zhang, Jian-Huang Lai, Zhanxiang Feng, and Xiaohua Xie. Uncertainty modeling with second-order transformer for group re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 3318–3325, 2022.

Quan Zhang, Jianhuang Lai, Zhanxiang Feng, and Xiaohua Xie. Uncertainty modeling for group re-identification. *International Journal of Computer Vision*, 132(8):3046–3066, 2024a.

Quan Zhang, Lei Wang, Vishal M Patel, Xiaohua Xie, and Jianhaung Lai. View-decoupled transformer for person re-identification under aerial-ground camera network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22000–22009, 2024b.

Xu Zhang, Zhiguang Wu, Qinghua Zhang, and Zuyu Zhang. Parallel branches-based second-order transformer for robust group re-identification with layout-guided occlusion mitigation. *Expert Systems with Applications*, pp. 128679, 2025.

Zhengyou Zhang. Perspective transformation. In *Computer Vision: A Reference Guide*, pp. 962–962. Springer, 2021.

Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR 2011*, pp. 649–656. IEEE, 2011.

Chen Zhou, Ming Han, Qi Liang, Yi-Fei Hu, and Shu-Guang Kuai. A social interaction field model accurately identifies static and dynamic social groupings. *Nature human behaviour*, 3(8):847–855, 2019.

Ji Zhu, Hua Yang, Weiyao Lin, Nian Liu, Jia Wang, and Wenjun Zhang. Group re-identification with group context graph neural networks. *IEEE Transactions on Multimedia*, 23:2614–2626, 2020.

A LIMITATIONS.

Althoug the proposed SIM method has achieved good results in methods and effects, it also exist some indubitable shortcomings. First, to reflect the members differentiation, the calculation of interaction probability requires skeleton extraction for orientation, which is restricted by the existing basic technology. Although we have adopted multiple models, the accuracy rate cannot reach perfection. Interaction probability is mainly determined by the distance, although the error caused by the orientation and openness is very small, the error still exist. Second, due to publicity and other reasons, the experiments was only conducted on the three most commonly used datasets.