Uncertainty Modeling with Second-Order Transformer for Group Re-identification

Quan Zhang[1], **Jian-Huang Lai**[1,2,3,4*], **Zhanxiang Feng**[1], **Xiaohua Xie**[1,2,3]
[1]School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou, China
[2]Guangdong Key Laboratory of Information Security Technology, Guangzhou, China
[3]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
[4]Key Laboratory of Video and Image Intelligent Analysis and Applicaiton Technology, Ministry of Public Security, China
zhangq48@mail2.sysu.edu.cn, {stsljh, fengzhx7, xiexiaoh6}@mail.sysu.edu.cn

Abstract

Group re-identification (G-ReID) focuses on associating the group images containing the same persons under different cameras. The key challenge of G-ReID is that all the cases of the intra-group member and layout variations are hard to exhaust. To this end, we propose a novel uncertainty modeling, which treats each image as a distribution depending on the current member and layout, then digs out potential group features through random sampling. Based on potential and original group features, uncertainty modeling can learn better decision boundaries, which is implemented by the member variation module (MVM) and layout variation module (LVM). Furthermore, we propose a novel second-order transformer framework (SOT), which is inspired by the fact that the position modeling in the transformer is coped with the G-ReID task. SOT is composed of the intra-member module and inter-member module. Specifically, the intra-member module extracts the first-order token for each member, and then the inter-member module learns a second-order token as a group feature by the above first-order tokens, which can be regarded as the token of tokens. A large number of experiments have been conducted on three available datasets, including CSG, DukeGroup and RoadGroup, which show that the proposed SOT outperforms all previous state-of-the-art methods.

Introduction

Group re-identification (G-ReID) aims to associate group images containing the same members under different cameras with non-overlapping views based on their similarity. G-ReID usually focuses on groups of $2 \sim 6$ members, and images belonging to the same group class should contain at least 60% same members. G-ReID is a more critical and challenging task than person re-id because people usually have group and social attributes, which indicates people prefer group moving in most real scenes. Therefore, G-ReID needs to deal with the member and layout variation. Specifically, the member variation means the number of intra-group members could decrease due to the member leaving or serve occlusion, and the layout variation means that the spatial positions may change under different cameras.

Although there are some pioneering works (Huang et al. 2021; Lin et al. 2021; Zhu et al. 2020; Yan et al. 2020) based on deep learning to address the above challenges, the performances are not satisfactory. The shortcomings are mainly due to the following two reasons. 1) The features extracted by the existing works are the specific features of the group image under the fixed member and layout. As shown in Fig. 1, The class boundaries learned from the orange triangles and green squares are the local representations of the whole classes, which leads to the fact that the decision boundary (red dotted line) based on the local boundaries cannot distinguish well between two classes. 2) The existing models are based on a combined framework of CNN and GNN, which are weak to describe the group layout feature due to the drawback of structure itself in position modeling, so the performances are limited.

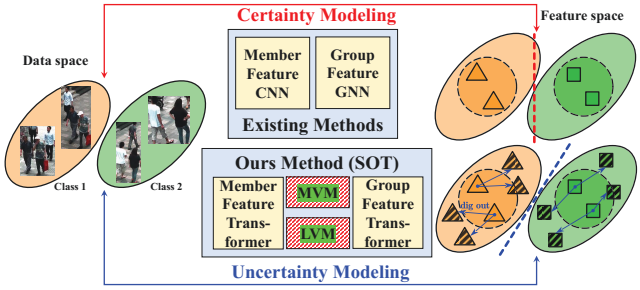


Figure 1: Certainty modeling versus uncertainty modeling. The pure triangles and squares represent the feature representations of the corresponding images. The textured triangles and squares represent potential group features mined from the original image through the MVM and LVM. The dotted circles represent the class boundaries learned from the given images by certainty modeling. The dotted lines represent the decision boundaries between the two classes.

In this paper, we propose a novel uncertainty modeling, which is motivated by the fact that variations of member and layout contained in each group are infinitely diverse. All situations cannot be exhausted no matter how elaborately sampling from the real world. Therefore, uncertainty is an inherent attribute of the group image that cannot disappear by collecting large-scale data. The proposed uncertainty modeling treats each group image as a distribution rather than a specific sample, and then digs out several potential group fea-

*Corresponding Author.

tures of the current group under other possible members and layouts by dynamically sampling on the distribution. Two modules, the member variation module (MVM) and layout variation module (LVM), are designed to construct the specific probability distribution for each image. As shown in Fig. 1, the group feature (triangles and squares) learned with uncertainty modeling are closer to the true boundary and consistent with the real-world distribution. Training and optimizing this true boundary can obtain more separable decision boundaries and more robust feature representations.

Specifically, the proposed MVM defines a random variable p to describe the probability distribution of intra-group member variation. A standard form p is constructed with the following properties. First, the group members tend to maintain stability when they occur across multi-cameras. Second, the variation probability will decrease with the increase of disappeared members when variation happens occasionally. Considering that the input does not always contain all the members of its group class, we will dynamically constrain the standard p to fit each image.

LVM focuses on the layout variation of each member. Because it is hard to exhaust all the spatial positions, LVM normalizes all possible spatial positions under a certain number of members into the same layout feature. To this end, a learnable memory bank M is designed to describe the layout features. For a group with j members, the j-th column of M is adopted as the normalized layout feature for each member. The advantage of LVM is that normalized layout can avoid oversampling on continuous position distribution.

Furthermore, we propose a **S**econd-**O**rder **T**ransformer model (SOT) inspired by the position embedding in transformer, which is coped with layout feature in G-ReID. The traditional CNN-GNN models lack spatial position modeling, which leads to low performance and can be overcome by our model. The proposed SOT consists of the intra-member and inter-member modules. For a group image, SOT crops each member firstly, and then partitions each member into several sub-patches. The intra-member module extracts the first-order token as each member feature by modeling the relationship among sub-patches through the member feature transformer. Then inter-member module models the group relationship among members through uncertainty modeling and extracts the second-order token as the group feature through the group feature transformer which receives the first-order tokens and outputs token of tokens.

Our contributions are summarized as follows.

- We propose the uncertainty modeling, which regards each image as a distribution instead of a specific sample. Uncertainty modeling aims to explore potential group variations through random sampling on distributions, which is achieved by the proposed member variation module (MVM) and layout variation module (LVM).
- We propose the second-order transformer (SOT), extracting the token as the member feature and the token of tokens as the group feature. SOT can efficiently extract the layout feature, which is hard in the existing methods.
- The SOT achieves the Rank-1/mAP of 91.7%/90.7%, 72.7%/78.9%, and 86.4%/91.3% on CSG, DukeGroup, and RoadGroup datasets, outperforming the state-of-the-art method by 28.5%, 15.3%, and 1.9% on Rank-1.

Related Work

Person Re-identification. Person re-identification (ReID) aims to associate individual pedestrians in a camera network with non-overlapping views. Recently, many methods (Sun et al. 2018; Wang et al. 2018; Dai et al. 2021; He et al. 2021b; Bai et al. 2021; Zhao et al. 2021; Wu, Zhu, and Gong 2022) based on deep learning have made significant progress in this field, including extracting more discriminative features and designing more suitable metrics. For example, OSNet (Zhou et al. 2019) and OSNet-AIN (Zhou et al. 2021) designed a novel backbone that both consider the discriminative feature learning and the computational cost. AGW (Ye et al. 2021) proposed a weighted regularization triplet metric learning method.

However, the above works were not suitable for G-ReID, because these work only focused on the appearance feature of the individual pedestrian, and ignored the relationship between intra-group members. The proposed SOT overcomes the shortcomings of existing work, and explicitly models the number and layout relationships of members, which greatly improves the performance.

Group Re-identification. Compared with ReID, G-ReID is less studied, and only a few pioneering works try to address this task. Some early works (Zheng, Gong, and Xiang 2009; Cai, Takala, and Pietikäinen 2010; Zhu, Chu, and Yu 2016; Lisanti et al. 2017) took the whole image as the input of the model, and directly extracted group features. Because these works were based on hand-crafted features, and background information was considered, the performance was not satisfactory. Recently, CNN-based works (Mei et al. 2020, 2019, 2021) have become the mainstream research, which cropped the intra-group members, and then extracted the group features. For example, DotGNN (Huang et al. 2019) adopted CycleGAN (Zhu et al. 2017) to obtain the style transfer, and then integrate the member features with GNN to extract group features. MRF (Xiao et al. 2018; Lin et al. 2021) considered more granular memberships, and proposed a multi-order matching method to calculate the similarity. GCGNN (Zhu et al. 2020) used K-nearest members to encode the each member, and then designed group context GNN to extract group features. MACG (Yan et al. 2020) proposed a multi-attention context graph framework which applied the complex attention mechanism to the group feature learning.

The performance of the above works are not satisfactory, mainly because: 1) They were based on the CNN and GNN framework, which were weak for modeling group layout; 2) They belonged to the certainty modeling. The proposed SOT can overcomes these shortcomings.

Transformer. Transformer (Vaswani et al. 2017) was proposed to extract text features in NLP task, and then generalized to many CV tasks and achieved good performances. For example, IPT (Chen et al. 2021) adopted the large-scale pre-training transformer to achieve good performance on many

low-level vision tasks. ViT (Dosovitskiy et al. 2021) is a pure transformer which directly divided the image into several patches. SwinTransformer (Liu et al. 2021) achieved a satisfied performance on object detection. DETR (Carion et al. 2020) proposed an end-to-end framework which combined the encoder and decoder together on object detection. TransReID (He et al. 2021a) first introduced the transformer into the person re-identification. However, transformer has not received too much attention in the G-ReID. To this end, we propose the second-order transformer to deal with G-ReID.

Method

In this section, we firstly introduce the MVM and LVM of uncertainty modeling, and then describe the proposed SOT network. Fig. 2 illustrates the method in detail.

Member Variation Module (MVM)

In this paper, MVM aims to construct a specific probability distribution for each image, and determines the existence of intra-group members by random sampling. Therefore, the key issue is how to obtain the specific form of probability distribution. We constrain the probability distribution to meet the following two properties, so that it can simulate the variations in the real-world scenes.

- **Stability**: For a robust group, the number of intra-group members usually remains unchanged.
- **Randomness**: When the robust group occasionally change, the probability of changing Z_d members will decrease significantly as Z_d increases.

Formally, the probability distribution can be described as $Pr\{p; Z_c, Z_t\}$, where Z_t and Z_c represent the number of members in the steady state and in the current image, and $Z_t = Z_c + Z_d$. We start with the trivial case $Z_c = Z_t$, and the symbol $Pr\{p; Z_t, Z_t\}$ can be abbreviated as $Pr\{p\}$. According to these two properties, the probability distribution function of p can be described as follows:

$$\begin{cases} Pr\left\{p = 0\right\} = P_0 \\ Pr\left\{0 < p \leqslant p_{\max}\right\} = \int_0^{p_{\max}} f\left(p\right) dp = 1 - P_0 \end{cases}, \quad (1)$$

where steady state probability $P_0 \in (0, 1)$ determines the probability that the group is in a stable state, cut-off probability $p_{max} \in (0, 0.4]$ determines the upper bound of the p, and $f(p)$ is the probability density function of p. Setting the upper bound of p_{max} to 0.4 is based on the member definition of G-ReID for the same group class.

Next, we derive the specific expression form of $f(p)$. We assume that the form of the $f(p)$ follows the truncated Gaussian distribution $\mathcal{N}(\mu, \sigma)$ which is satisfied with "randomness" of p. Due to the sampling space $(-\infty, +\infty)$ of the $\mathcal{N}(\mu, \sigma)$ is not consistent with the sampling interval $[0, p_{\max}]$ of p, we impose the following two constraints on the $\mathcal{N}(\mu, \sigma)$. First, the probability of the $\mathcal{N}(\mu, \sigma)$ sampling in the interval $(-\infty, 0)$ is mapped to the probability when $p = 0$. Second, the $\mu+3\sigma$ in Gaussian distribution is mapped to $p_{\max}$, which ensures that $Pr\{p \in (p_{\max}, +\infty)\}$ is a small probability event. So far, p follows the conditional Gaussian distribution $\mathcal{N}(\mu, \sigma; P_0, p_{\max})$ under the tolerable error.

After that, the solution of μ and σ can be obtained through these two constrains, which can be described as follows:

$$\begin{cases} \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(p-\mu)^2}{2\sigma^2}} dp = P_0 \\ \mu + 3\sigma = p_{\max} \end{cases}. \quad (2)$$

Solving the Eq. 2, we can get the followings:

$$\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(p-\mu)^2}{2\sigma^2}} dp = \frac{1}{2}\left[1 - erf\left(\frac{\mu}{\sqrt{2}\sigma}\right)\right], \quad (3)$$

where $erf\left(x\right) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$ is Gauss error function, and $erf(x) \in [-1, 1]$. Solving the Eq. 3, we can get the followings:

$$\frac{\mu}{\sqrt{2}\sigma} = erf^{-1}\left(1 - 2P_0\right). \quad (4)$$

From the Eq. 1, we can see the $P_0 \in (0, 1)$. Therefore, Eq. 4 is always satisfied because the interval of $1 - 2P_0$ is included in the definition domain of $erf^{-1}(\cdot)$. After that, the analytical solutions for the μ and σ are described as follows:

$$\begin{cases} \sigma = \frac{p_{\max}}{\sqrt{2}erf^{-1}(1-2P_0)+3} \\ \mu = p_{\max} - 3\sigma \end{cases}. \quad (5)$$

The non-trivial case $Z_c < Z_t$ is an extension of the above conclusion. The upper bound of the member variation needs to be corrected in order to meet the definition of G-ReID, which can be described and solved as follows:

$$Z_c - Z_c p_{\max}^{'} \geqslant Z_t - Z_t p_{\max} \quad (6)$$

$$\Longrightarrow p_{\max}^{'} = \max\left(0, 1 - \left(1 - p_{\max}\right) Z_t / Z_c\right), \quad (7)$$

where $p_{\max}^{'}$ stands for the true cut-off probability under the current image. In the non-trivial case, we use P_0 and $p_{\max}^{'}$ to solve the μ and σ via Eq. 5. In a training batch consisting of several group images, we count the largest number of members contained in the current class as Z_t. Each member is determined whether to change by the same p sampled on the $\mathcal{N}(\mu, \sigma; P_0, p_{\max}^{'})$, and the remaining members are then modeled for layout features and group feature extraction.

Layout Variation Module (LVM)

Compared with MVM, the modeling of the layout is a more challenging problem. Because the spatial position variation of each member is a continuous distribution in the image, which is hard to exhaust all cases. To this end, we propose a normalized layout representation to avoid the above infinite enumerations. Specifically, a learnable layout embedding bank $M \in R^{D \times M_0}$ is designed, and each column of M represents the layout feature under a certain member number, where the D is the layout feature dimension and M_0 is the maximum number of group members in the training set. For a group with j remaining members after MVM, we select the j-th column in M as the layout feature of each member in the current group, which is shown in Fig. 2.

In the testing stage, if the members exceed M_0, a random D-dim vector is used as the current layout feature. Because the groups with more than M_0 members are not satisfied with the definition of G-ReID, and are regarded as the distractors that needs to be discarded. Using a random layout feature can effectively reduce the similarity with other groups, and avoid the wrong matching.

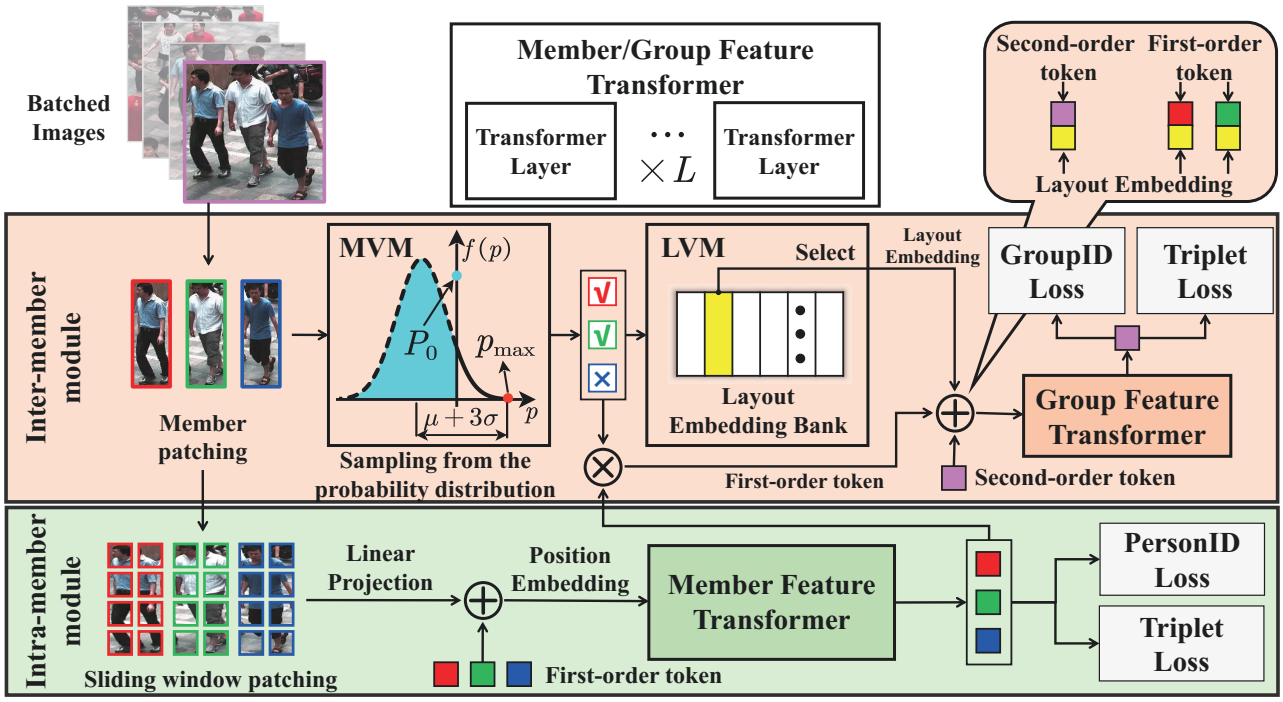


Figure 2: The illustrate of the proposed SOT. The pink square stands for the group feature, and the red, green and dark blue square represent the member features. Group and member feature transformer are consisted of transform layers with different L, and do not share parameters. ✓/× from MVM stands for the existence/disappearance of the current member.

Second-Order Transformer (SOT)

The whole structure of the SOT is shown in Fig. 2. The SOT is composed of the intra-member and inter-member modules. The intra-member module considers the feature extraction of individual members in the group, and the inter-member module focuses on the uncertainty modeling of group variation and the group feature extraction. We adopt a transformer module to implement the extraction of member and group features, which is composed of different numbers of transformer layers. Notably, the member and group feature transformers do not share parameters.

As shown in Fig. 2, for a group image in the batch with N images, we first crop the region of each member according to the ground truth. Then, we send each member patch to the intra-member module and divide it into fixed-size image sub-patches (specifically, 16×16), and add a first-order token and sequential position embeddings to these sub-patches. The first-order token is regarded as the member feature after passing through the member feature transformer, which is supervised by the person identity and triplet loss function.

$$\mathcal{L}_{ID} = -\frac{1}{P}\sum_{j=1}^{P}\sum_{i=1}^{C} y_{ji}\log\left(\widehat{y}_{ji}\right), \tag{8}$$

where P represents the total member number of the current batch, C represents the total member classes, the indicator function $y_{ji} = \mathbb{1}(j = i)$ 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.

$$\mathcal{L}_{Tri} = \frac{1}{P}\sum_{i=1}^{P}\max\left(d\left(f_i, f_i^+\right) - d\left(f_i, f_i^-\right) + m, 0\right), \tag{9}$$

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.

$$\mathcal{L}_p = \mathcal{L}_{ID} + \mathcal{L}_{Tri} \tag{10}$$

In the inter-member module, MVM dynamically samples probability values from the proposed probability distribution in MVM for the current group, and then determines whether the first-order token of each member is discarded. After that, LVM selects the corresponding column in the layout embedding bank as the layout feature of each member according to the number of remaining members. We add the second-order token as the token of these first-order tokens, and the second-order token can extract the group feature through the group feature transformer. The loss function $\mathcal{L}_g$ of a second-order token 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 of the SOT is described as follows:

$$\mathcal{L} = \mathcal{L}_p + \mathcal{L}_g. \tag{11}$$

It can be seen that in the training phase of the SOT, MVM and LVM will dynamically change the members and layout representation of the current group, in order to mine more potential feature representations.

Method	Publication	CSG				DukeGroup				RoadGroup			
		Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
CRRRO-BRO	BMVC 2009	10.4	25.8	37.5	-	9.9	26.1	40.2	-	17.8	34.6	48.1	-
Covariance	ICPR 2010	16.5	34.1	47.9	-	21.3	43.6	60.4	-	38.0	61.0	73.1	-
PREF	ICCV 2017	19.2	36.4	51.8	-	30.6	55.3	67.0	-	43.0	68.7	77.9	-
BSC+CM	ICIP 2016	24.6	38.5	55.1	-	23.1	44.3	56.4	-	58.6	80.6	87.4	-
LIMI	MM 2018	-	-	-	-	47.4	68.1	77.3	-	72.3	90.6	94.1	-
DOTGNN	MM 2019	-	-	-	-	53.4	72.7	80.7	-	74.1	90.1	92.6	-
GCGNN	TMM 2020	-	-	-	-	53.6	77.0	<u>91.4</u>	-	81.7	94.3	96.5	-
MGR	TCYB 2021	57.8	71.6	76.5	-	48.4	75.2	89.9	-	80.2	93.8	96.3	-
MACG	TPAMI 2020	<u>63.2</u>	<u>75.4</u>	<u>79.7</u>	-	<u>57.4</u>	<u>79.0</u>	90.3	-	<u>84.5</u>	<u>95.0</u>	<u>96.9</u>	-
SOT (Ours)	-	**91.7**	**96.5**	**97.6**	**90.7**	**72.7**	**88.6**	**93.2**	**78.9**	**86.4**	**96.3**	**98.8**	**91.3**

Table 1: Comparison of the proposed method with the state-of-the-arts on CSG, DukeGroup and RoadGroup. The compared methods are categorized into two groups, including hand-crafted methods and deep learning methods. The best and second best results are shown in bold and underline respectively. The Rank-1, Rank-5, Rank-10 and mAP are reported(%).

Experiments

Datasets

The proposed SOT is evaluated on DukeGroup (Lin et al. 2021), RoadGroup (Lin et al. 2021) and CSG (Yan et al. 2020) datasets. The DukeGroup dataset contains 354 images including 177 group classes. The RoadGroup dataset contains 324 images including 162 group classes. Follow the protocol in (Lin et al. 2021), the training and testing set of DukeGroup and RoadGroup are randomly and equally split.

The CSG dataset contains 3,839 images including 1,558 group classes, where 859/699 groups are split for training/testing. Follow the protocol in (Yan et al. 2020), the images in the test set are sequentially selected as the probe, and all the remaining images are regarded as the gallery. In addition, CSG adds extra 5K group images as distractors in the gallery. We do not use any extra datasets when training on each G-ReID dataset for fair performance comparison. The Cumulative Matching Characteristics (CMC) at Rank-1, Rank-5, Rank-10, and mean Average Precision (mAP) are used as evaluation metrics.

Details

We adopt ViT-Base (Dosovitskiy et al. 2021), pre-trained on ImageNet (Deng et al. 2009), as the backbone of the member feature transformer. We regard the SOT without LVM and MVM as the certainty modeling. For the input group image, we crop all the member patches by the given bounding box and resize them to 256×128. In the training stage, the random horizontal flip and random erasing are performed with a fixed probability of 0.5. Each mini-batch is sampled with 16 group identities, and each group identity selects 4 images. We choose SGD (Bottou 2012) as the optimizer. Our training stage ends when the iteration number reaches 400 epochs. We use a cosine annealing learning rate strategy. The initial learning rate is 2e-3, and the minimum learning rate is 1.6e-4. The learning rate of the inter-member module is multiplied by 0.1. The weight decay is 1e-4. The selection of hard samples in triplet loss adopts an online mining strategy. In the testing stage, we do not use any data augmentation and re-ranking. We use the Euclidean distance to measure the normalized features. All ablation studies, parameter analyses, and visualizations have been conducted on the DukeGroup dataset if there is no additional comments.

Performance

We evaluate the proposed SOT method against the existing methods on three available G-ReID datasets to show the superiority of our method. As shown in Table 1, we divide the existing methods into two groups: hand-crafted G-ReID methods including CRRRO-BRO (Zheng, Gong, and Xiang 2009), Covariance (Cai, Takala, and Pietikäinen 2010), PREF (Lisanti et al. 2017) and BSC+CM (Zhu, Chu, and Yu 2016); deep learning G-ReID methods including LIMI (Xiao et al. 2018), DOTGNN (Huang et al. 2019), GCGNN (Zhu et al. 2020), MGR (Lin et al. 2021) and MACG (Yan et al. 2020). The MACG is regarded as the state-of-the-art method in the existing methods according to the performance. Three conclusions can be drawn from Table 1.

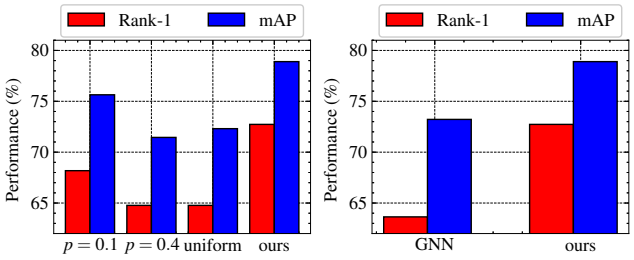
First, our SOT achieves very strong performance on CSG, DukeGroup and RoadGroup datasets, which far exceeds the MACG on Rank-1 and mAP. In the CSG datasets, the performance of our SOT achieves 91.7%/90.7% on Rank-1/mAP, and exceeds MACG by 28.5% on Rank-1. In the DukeGroup datasets, the performance of our SOT achieves 72.7%/78.9% on Rank-1/mAP, and exceeds MACG by 15.3% on Rank-1. In the RoadGroup datasets, the performance of our SOT achieves 86.4%/91.3% on Rank-1/mAP, and exceeds MACG by 1.9% on Rank-1. This shows that the SOT brings different degrees of performance gain on all datasets, proving that uncertainty is an attribute of group images and does not disappear as the data size increases. This also shows that the SOT overcomes the inherent uncertainty of group images and brings significant improvements.

Second, the performances of the methods based on the hand-crafted features are relatively low. Different from these works, the proposed SOT crops each member in the group to avoid background interference. Furthermore, the SOT designs a second-order transformer to extract the feature of each member and the whole group, which is more robust and discriminative.

Finally, the performances of deep learning methods are still unsatisfactory, which is caused by the certainty modeling and insufficient layout modeling. Different from these

MVM	LVM	CSG		DukeGroup		RoadGroup	
		Rank1	mAP	Rank1	mAP	Rank1	mAP
		85.56	84.40	65.91	75.00	83.95	89.10
✓		88.92	87.31	67.05	75.10	85.19	89.89
	✓	90.26	88.92	68.18	75.55	85.19	89.70
✓	✓	91.70	90.70	72.73	78.90	86.40	91.30

Table 2: Ablation study of the proposed SOT (%).



(a) Probability distribution. (b) Inter-member Module.

Figure 3: Comparisons with alternative variants.

works, the SOT designs the uncertainty modeling which mines potential group features by treating each image as a specific distribution. In addition, the proposed SOT leverages the transformer-based network to extract the member and group features, which is more suitable for G-ReID.

Ablation Study

Effect of MVM and LVM. The ablation experiment mainly shows the effect of two proposed modules, MVM and LVM, on uncertainty modeling. We mainly analyze the results on the DukeGroup, and there are similar conclusions on the other two datasets. As shown in Table 2, two conclusions can be drawn. First, each module can improve the performance when used alone. Compared with the baseline model, MVM increased by +1.14%/+0.1% on Rank-1/mAP, and LVM increased by +2.27%/+0.55% on Rank-1/mAP. This shows that each module digs for potential member variation and layout variation, respectively, making SOT more discriminative. Second, when two modules are both used, the performance gain, +6.82%/+3.90% on Rank-1/mAP, is higher than than the sum of individual modules. This shows that member and layout variations are two complementary aspects for modeling group variations. Using MVM and LVM simultaneously can explore more potential group features.

Alternative Probability Distribution. In order to verify the effectiveness of the proposed probability distribution $\mathcal{N}(\mu, \sigma; P_0, p'_{\max})$, we choose other distributions as comparisons, including fixed probability $p = 0.1$ and $p = 0.4$, and uniform distribution in the interval $[0, 0.4]$. As shown in Fig. 3(a), our distribution is better than other alternative distributions. The proposed distribution is coped with the real scene, due to the constraints of "stability" and "randomness". In addition, we construct a specific distribution for each image, which alleviates the class confusion caused by the change of members.

Layout modeling strategy	Rank-1	mAP
no embedding	67.05	75.04
random sequential embedding	68.18	76.62
normalized embedding (Ours)	**72.73**	**78.90**

Table 3: Several alternative layout feature modeling (%).

Alternative Inter-member Module. To verify that group feature transformer is more suitable for G-ReID, we select a classical GNN (Hamilton, Ying, and Leskovec 2017) model for comparison. As shown in Fig. 3(b), the existing GNN lacks the extraction of member layout information and only learns group features through member appearance features, which is not sufficient and robust. On the contrary, our method can model layout features when extract group features, thus, it can bring more performance gain and is more suitable for G-ReID than GNN.

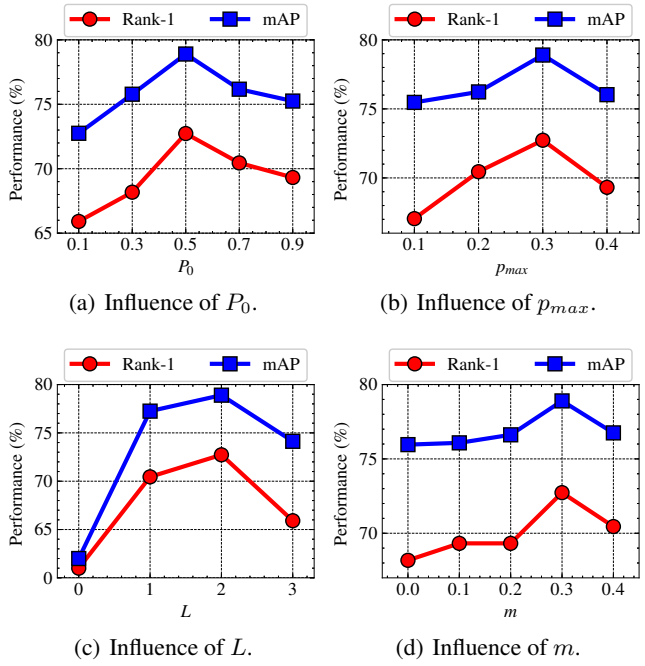
Alternative Layout Feature Modeling. We design several other layout feature modeling strategies and compare them with our method. No embedding means that the layout embedding bank in LVM is discarded, but we extract group features through member appearance features. Random sequential embedding is similar to the position embedding in ViT, which means that we assign group members a random sequence to enumerate possible layout situations. As shown in Table 3, row 1 not only ignores layout variations but also ignores the discrimination information contained in layout features themselves. Row 2 proves that the layout variations cannot be exhausted by finite enumeration, so the performance is also limited. Our method brings the best performance when describing the layout uncertainty.

Parameter Analysis

Influence of P_0. P_0 controls the effect of stability in MVM, which determines that each input maintains the current members and layout by probability P_0 during training. When P_0 is too small, the stability of the SOT cannot be guaranteed, which lead to the performance degradation; when P_0 is too large, the SOT is too stable to ignore the mining of potential features, which lead to insufficient model generalization and performance degradation. As shown in Fig. 4(a), we set $P_0 = 0.5$ for the best performance.

Influence of p_{max}. p_{max} controls the effect of randomness in MVM, which controls the maximum variation probability of each member in the current group during training, and also reflects the drastic degree of member variation. As shown in Fig. 4(a), the drastic degree of member variation becomes greater with the increase of p_{max}, and the potential group features mined are more and more diverse. When p_{max} reaches 0.3, the model achieves the best performance, so we set $p_{max} = 0.3$. When p_{max} further increases, some unexpected variations that cause confusion of group class will happen, which lead to the performance degradation.

Influence of L. We analyze the relationship between the number of layers in the group feature transformer and performance and show the results in Fig. 4(c). When $L = 0$, there is no connection between the second-order token and



(a) Influence of P_0. (b) Influence of p_{max}.

(c) Influence of L. (d) Influence of m.

Figure 4: Parameter analysis of the proposed SOT.

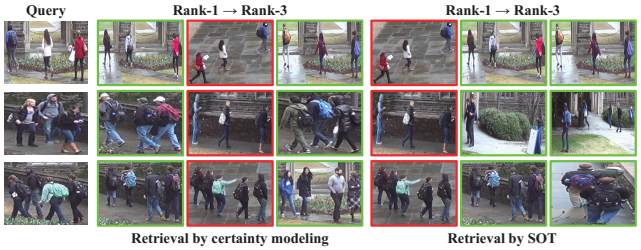


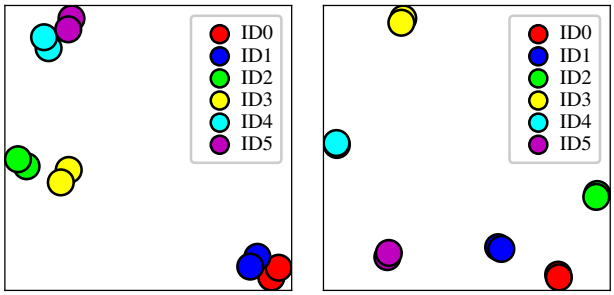
Figure 5: The top-three retrieval visualization of the certainty modeling and the SOT. The red/green boxes indicate the correct/wrong matches. In the DukeGroup dataset, each query only has one correct matching in the gallery.

member features. Therefore, the performance is very low. With the increase of L, the performance starts to improve and achieves the best with $L = 2$. When L further increases to 3, there are too many parameters resulting in over-fitting, which reduces the performance.

Influence of m. Margin m controls both intra-member/group-class consistency and inter-member/group-class discrepancy. We select five different values $\{0, 0.1, 0.2, 0.3, 0.4\}$ to analyze the effect of m on the performance. As shown in Fig. 4(d), we set $m = 0.3$ as the best performance of the SOT.

Visualization

Retrieval visualization. Fig. 5 shows the top-three retrieval visualization of the certainty modeling and the proposed SOT. The advantages of SOT are reflected in the following two aspects. 1) Certainty modeling tends to search for images with similar layouts and cannot model layout variations. For example, certainty modeling tends to search for images with a horizontal layout in row 1 and images with a tight layout in row 2. However, SOT can get the correct results under the large layout differences between query and gallery. 2) Certainty modeling tends to search for the images with the same number with query. For example, certainty modeling tends to search for images with four members in row 3. However, SOT extracts similar group features from the images with different members of the same group.



(a) Certainty modeling. (b) SOT.

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

Feature visualization. Fig. 6 shows the feature distribution visualization of six classes in the training set when the certainty modeling and the proposed SOT are trained to converge. It is worth noting that there are only two images in each group class. In Fig. 6(a), certainty modeling learns poor decision boundaries (cyan and purple circles, blue and red circles) due to the lack of potential group features mining. With the advantages of uncertainty modeling, the feature distribution of the SOT shows obvious intra-class consistency and inter-class discrepancy.

Conclusion

In this paper, we focus on member and layout variation in G-ReID. To this end, firstly, we propose a novel uncertainty method to model the variation of intra-group members and layout. The advantage of uncertainty modeling is that lots of potential group features can be explored and the training of the model can be promoted. Secondly, we propose a second-order transformer (SOT) to extract the features of individual members and groups, respectively. Finally, the proposed SOT achieves the state-of-the-art performance on multiple datasets, which greatly exceeds the existing methods.

Acknowledgments

This project was supported by the NSFC (62076258), the Project of Natural Resources Department of Guangdong Province ([2021]34), and the Project of Ministry of Public Security of China (2019GABJC39).

References

Bai, Y.; Jiao, J.; Ce, W.; Liu, J.; Lou, Y.; Feng, X.; and Duan, L.-Y. 2021. Person30K: A Dual-Meta Generalization Network for Person Re-Identification. In *CVPR*, 2123–2132.

Bottou, L. 2012. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, 421–436.

Cai, Y.; Takala, V.; and Pietikäinen, M. 2010. Matching Groups of People by Covariance Descriptor. In *ICPR*, 2744–2747.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *ECCV*, 213–229.

Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; and Gao, W. 2021. Pre-Trained Image Processing Transformer. In *CVPR*, 12299–12310.

Dai, Y.; Li, X.; Liu, J.; Tong, Z.; and Duan, L.-Y. 2021. Generalizable Person Re-Identification With Relevance-Aware Mixture of Experts. In *CVPR*, 16145–16154.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Li, F. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive Representation Learning on Large Graphs. In *NeurIPS*, volume 30.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021a. TransReID: Transformer-Based Object Re-Identification. In *ICCV*, 15013–15022.

He, T.; Shen, X.; Huang, J.; Chen, Z.; and Hua, X.-S. 2021b. Partial Person Re-Identification With Part-Part Correspondence Learning. In *CVPR*, 9105–9115.

Huang, Z.; Wang, Z.; Hu, W.; Lin, C.; and Satoh, S. 2019. DoT-GNN: Domain-Transferred Graph Neural Network for Group Re-identification. In *ACMMM*, 1888–1896.

Huang, Z.; Wang, Z.; Tsai, C.-C.; Satoh, S.; and Lin, C.-W. 2021. DotSCN: Group Re-Identification via Domain-Transferred Single and Couple Representation Learning. *IEEE TCSVT*, 31(7): 2739–2750.

Lin, W.; Li, Y.; Xiao, H.; See, J.; Zou, J.; Xiong, H.; Wang, J.; and Mei, T. 2021. Group Reidentification with Multi-grained Matching and Integration. *IEEE TCYB*, 51(3): 1478–1492.

Lisanti, G.; Martinel, N.; Bimbo, A. D.; and Foresti, G. L. 2017. Group Re-identification via Unsupervised Transfer of Sparse Features Encoding. In *ICCV*, 2468–2477.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *ICCV*, 10012–10022.

Mei, L.; Lai, J.; Feng, Z.; and Xie, X. 2020. From pedestrian to group retrieval via siamese network and correlation. *Neurocomputing*, 412: 447–460.

Mei, L.; Lai, J.; Feng, Z.; and Xie, X. 2021. Open-World Group Retrieval with Ambiguity Removal: A Benchmark. In *ICPR*, 584–591. IEEE.

Mei, L.; Lai, J.; Xie, X.; Zhu, J.; and Chen, J. 2019. Illumination-invariance optical flow estimation using weighted regularization transform. *IEEE TCSVT*, 30(2): 495–508.

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018. Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline). In *ECCV*, 501–518.

van der Maaten, L.; and Hinton, G. 2008. Visualizing Data using t-SNE. *JMLR*, 9(86): 2579–2605.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In *NeurIPS*, volume 30.

Wang, G.; Yuan, Y.; Chen, X.; Li, J.; and Zhou, X. 2018. Learning Discriminative Features with Multiple Granularities for Person Re-Identification. In *ACMMM*, 274–282.

Wu, G.; Zhu, X.; and Gong, S. 2022. Learning hybrid ranking representation for person re-identification. *PR*, 121: 108239.

Xiao, H.; Lin, W.; Sheng, B.; Lu, K.; Yan, J.; Wang, J.; Ding, E.; Zhang, Y.; and Xiong, H. 2018. Group Re-Identification: Leveraging and Integrating Multi-Grain Information. In *ACMMM*, 192–200.

Yan, Y.; Qin, J.; Ni, B.; Chen, J.; Liu, L.; Zhu, F.; Zheng, W.-S.; Yang, X.; and Shao, L. 2020. Learning Multi-Attention Context Graph for Group-Based Re-Identification. *IEEE TPAMI*, 1–1.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. 2021. Deep Learning for Person Re-identification: A Survey and Outlook. *IEEE TPAMI*.

Zhao, Y.; Zhong, Z.; Yang, F.; Luo, Z.; Lin, Y.; Li, S.; and Sebe, N. 2021. Learning to Generalize Unseen Domains via Memory-based Multi-Source Meta-Learning for Person Re-Identification. In *CVPR*, 6277–6286.

Zheng, W.; Gong, S.; and Xiang, T. 2009. Associating Groups of People. In *BMVC*, 1–11.

Zhou, K.; Yang, Y.; Cavallaro, A.; and Xiang, T. 2019. Omni-Scale Feature Learning for Person Re-Identification. In *ICCV*, 3701–3711.

Zhou, K.; Yang, Y.; Cavallaro, A.; and Xiang, T. 2021. Learning Generalisable Omni-Scale Representations for Person Re-Identification. *IEEE TPAMI*.

Zhu, F.; Chu, Q.; and Yu, N. 2016. Consistent matching based on boosted salience channels for group re-identification. In *ICIP*, 4279–4283.

Zhu, J.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *ICCV*, 2242–2251.

Zhu, J.; Yang, H.; Lin, W.; Liu, N.; Wang, J.; and Zhang, W. 2020. Group Re-identification with Group Context Graph Neural Networks. *IEEE TMM*, 1–1.