Modeling 3D Layout For Group Re-Identification

Quan Zhang[1], Kaiheng Dang[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, dangkh}@mail2.sysu.edu.cn, {stsljh, fengzhx7, xiexiaoh6}@mail.sysu.edu.cn

Abstract

Group re-identification (GReID) attempts to correctly associate groups with the same members under different cameras. The main challenge is how to resist the membership and layout variations. Existing works attempt to incorporate layout modeling on the basis of appearance features to achieve robust group representations. However, layout ambiguity is introduced because these methods only consider the 2D layout on the imaging plane. In this paper, we overcome the above limitations by 3D layout modeling. Specifically, we propose a novel 3D transformer (3DT) that reconstructs the relative 3D layout relationship among members, then applies sampling and quantification to preset a series of layout tokens along three dimensions, and selects the corresponding tokens as layout features for each member. Furthermore, we build a synthetic GReID dataset, City1M, including 1.84M images, 45K persons and 11.5K groups with 3D annotations to alleviate data shortages and poor annotations. To the best of our knowledge, 3DT is the first work to address GReID with 3D perspective, and the City1M is the currently largest dataset. Several experiments show the superiority of our 3DT and City1M. Our project has been released on https://github.com/LinlyAC/City1M-dataset.

1. Introduction

Group re-identification (GReID) aims to match groups with the same members under different cameras. Usually, we deal with groups of 2 to 6 members, and we treat group images with more than 60% of the same members as the same group class. GReID aims to bring positive services and contributions to human society and eliminate potential social risks, such as child trafficking and kidnapping.

*Corresponding Author.

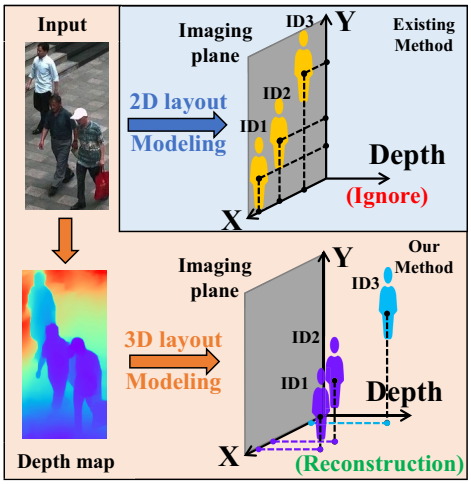


Figure 1. The illustration of our novelty. The X-Y plane represents the imaging plane. The Depth dimension represents the distance from person to camera. In the depth map, the darker the color, the closer to the camera, and vice versa.

GReID has potential applications in detecting and preventing these events, which protects the safety of citizens. The challenge of GReID is how to jointly model the appearance and layout features of group images.

Most existing methods [4, 9, 22, 23] adopt only the appearance features of groups. However, the appearances of group images are vulnerable to member occlusion and variations, leading to a large performance drop. In addition, some methods [24] attempt to extract features from layout relationships to alleviate the lack of appearances. Unfortunately, existing layout-based methods belong to the 2D modeling, which ignores member depth information and leads to unsatisfactory performance. We denote this shortcoming as the 2D layout ambiguity. As shown in Fig. 1, ID2 and ID3 are incorrectly modeled as neighbors on 2D

images, but they are far away in 3D scenes, which means that the real layout is hardly reconstructed without depth.

In this paper, we model the layout relationship from a 3D perspective, which can effectively eliminate the 2D layout ambiguity. Specifically, we calculate the depth of each member in the group image via depth estimation to reconstruct the 3D layout relationship of the group. As shown in Fig. 1, although ID2 and ID3 are adjacent in the X-Y plane, the depth map indicates that they have different depth components. Our method can correctly reflect this cue and reconstruct a relatively accurate group layout. Based on this, we propose a 3D Transformer (3DT), which performs sampling and quantization in the X-Y-D space, and presets a series of layout tokens along each dimension. 3DT calculates the average center position of each member and concatenates the corresponding layout's tokens in the three axes as the layout feature. Finally, 3DT extracts the group feature by joint modeling the appearance and layout features.

Furthermore, we find that existing datasets do not provide 3D labels, and constructing a dataset with rich labeling is highly expensive. Therefore, we contribute a synthetic GReID dataset, named *City1M*, which has the following three advantages. 1) Larger data scale. City1M includes 1.84M images, 45K persons and 11.5K groups. Compared with the current largest dataset CSG [19], the number of images and group identities are 600 times and 7 times that of CSG respectively. 2) More diversified samples. To simulate the real-world monitoring scene, City1M considers illumination variations, occlusions, resolution variations, intra-group member and layout variations. 3) More detailed annotations. Not only do we provide 3D position label of each member, but we also provide other annotations such as the shooting time, camera coordinates, and orientation. These advantages greatly facilitate the research of GReID.

Our contributions can be summarized as follows:

1. We propose the 3D Transformer (3DT) to perform 3D layout modeling, which eliminates the layout ambiguity in existing methods. To the best of our knowledge, we are the first 3D-based method. Compared with the 2D-based methods, our method can obtain more accurate layout features.

2. We propose a large-scale synthetic GReID dataset to alleviate data shortages and poor annotations, which contains 1.84M images with 11.5K groups and is three orders of magnitude larger than the existing dataset.

3. Lots of experiments demonstrate the superiority of the proposed 3DT and City1M. 3DT exceeds the existing methods by 29.7%, 25.6% and 6.9% on Rank1 on CSG, DukeGroup and RoadGroup. The 3DT+, pretrained on City1M, will further improve 2.2%, 7.9% and 2.4% on Rank1. Surprisingly, a strong performance has been achieved by testing the pretrained model directly on real datasets.

2. Related Work

2.1. Group Re-Identification

The deep learning methods of GReID surpassed the traditional methods [4, 11, 22, 23] and became mainstream, which are mainly divided into two aspects: appearance-based and layout-based methods. LIMI [18] and MGR [10] designed a multi-order network for multi-grain representations of groups. DotSCN [9] extracted the group consistency feature by learning the difference features of the pair members in the two images. DotGNN [8] adopted a graph convolution network to integrate the appearance group features. MACG [19] designed complex multiple attentions to capture the key group features.

These works focus on the appearance features and ignore layout features. GCGNN [24] calculated the spatial relationship among members to mine neighbors for enhancement. However, GCGNN only focuses on the layout relationship on the 2D image coordinates and ignores the ambiguity caused by the imaging process. In this paper, we focus on this limitation and propose a 3D-based layout relationship modeling method, which effectively alleviates the ambiguity of 2D layout.

2.2. Synthesized datasets

Synthesized datasets are also important for ReID, which is a low-cost and proven efficient approach. PersonX [12] introduced a synthetic data engine based on Unity3D [13], composed of hand-crafted 3D person models. RandPerson [15] proposed a method of combining UV maps with random colors and textures, generalizing quantities of person models with MakeHuman [3]. UnrealPerson [20] designed a low-cost pipeline to construct ReID datasets, and the synthesized images are more diverse and realistic. However, synthetic data is lacking attention in GReID. Our City1M is the first large-scale synthetic dataset in GReID.

2.3. Transformer

Transformer [14] was first proposed in NLP task, and then generalized to many CV tasks and achieved good performances. For example, TransReID [7] was the first work which introduced the transformer into the person ReID. However, transformer has not been raised too much attention in the GReID. To this end, we propose the PST because the position modeling of transformer is suitable for the layout modeling in the GReID.

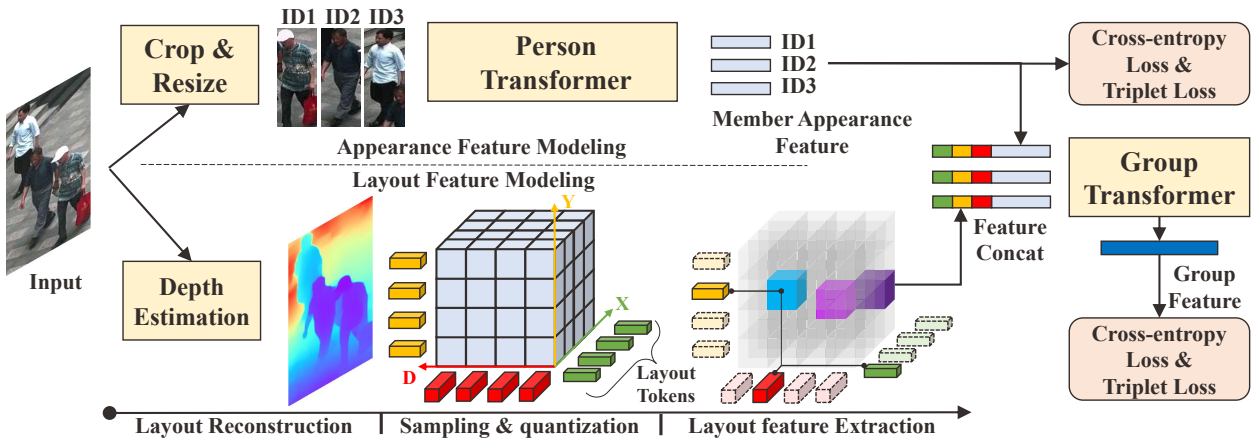


Figure 2. The illustration of our whole framework. The proposed 3D Transformer includes person transformer and group transformer with 3D layout tokens. The layout relationship is reconstructed in X-Y-D space. The X-Y plane represents the imaging plane. The "D" is depth dimension which represents the distance from person to camera. The green/yellow/red blocks are layout tokens which represents the corresponding position's feature. The colorful cubes stand for the member average center position in X-Y-D space.

3. Method

3.1. 3D Layout Reconstruction

Due to the existing datasets usually providing the two-dimensional position coordinates of members, an intuitive idea to reconstruct a 3D layout from 2D images is to estimate the depth of each member by using the depth estimation method. It should be emphasized that we do not need to reconstruct the accurate absolute depth information of the whole scene. The relative depth information among intra-group members is very sufficient for reconstructing the layout relationship. For a group image $\mathcal{I} \in \mathbb{R}^{3\times H\times W}$, we adopt a strong depth estimation method, Adabins [1], to obtain the depth map, which can be described as follows.

$$\mathcal{D} = N_{Ada}(\mathcal{I}), \tag{1}$$

where the depth map $\mathcal{D} \in \mathbb{R}^{H\times W}$ has the same size with $\mathcal{I}$, and each pixel of $\mathcal{D}$ is in the interval $(0, 1)$ and represents the relative depth. It should be noted that if the dataset provides depth information, such as the proposed City1M, the above estimation process can be omitted.

After that, we define the depth of each member as the average of all pixels in a given bounding box of $\mathcal{D}$, which can be described as follows.

$$\mathcal{D}_i = \frac{1}{h_i \times w_i} \sum_{s=0}^{h_i-1} \sum_{t=0}^{w_i-1} \mathcal{D}(x_i + s, y_i + t), \tag{2}$$

where $\mathcal{D}_i$ represents the average depth of the i-th member, and $Rect[x_i, y_i, h_i, w_i]$ represents the given bounding box of the i-th member.

Similarly, we use the normalized center coordinate of the bounding box as the two-dimensional relative position. The coordinates P_i of i-th member can be described as follows.

$$P_i = \left(\frac{(x_i + h_i/2)}{H}, \frac{(y_i + w_i/2)}{W}, \mathcal{D}_i \right). \tag{3}$$

3.2. 3D Transformer

The whole structure of our 3D Transformer (3DT) is illustrated in Fig. 2. The proposed 3DT mainly consists of the person transformer and group transformer with 3D layout tokens. For a group image, we first crop and resize each member, and send them to a person transformer network, such as ViT [6], to extract the appearance features. Next, we model the layout relationship among members from the original image and extract the layout features of each member. Then, we concatenate the appearance features of each member with the corresponding layout features. Finally, we integrate all members features in the group transformer to obtain the feature representation of the whole group.

In the above processes of 3DT, the core step is to extract each member's layout features from the layout relationship obtained in Sec. 3.1. 3D layout reconstruction essentially constructs a normalized continuous X-Y-D space. However, the position coordinates in X-Y-D space are inexhaustible, which leads to difficult layout feature extraction. Therefore, we perform sampling and quantization operations on the X-Y-D space and convert it into a discrete space. The sampling operation means that the X-Y-D space is evenly divided along three dimensions by the sampling rate $\frac{1}{\sigma}$, which divides each dimension into σ blocks and divides the original space into σ^3 cubes. For example, the $\frac{1}{\sigma}$ is set to $\frac{1}{4}$, and the original space is divided into 4^3 cubes.

Quantization operation means that σ blocks in each dimension are assigned with σ learnable feature embeddings,

Table 1. Detailed comparisons of mainstream datasets. "P/G" stands for the traditional person ReID and group ReID. 3D position label stands for providing spatial coordinates for each person in three dimensions. Camera orientation stands for providing the position and view of each camera in the 3D scene.

Dataset	Task	Real or Synthetic	#Images	#Cameras	#PersonID	#GroupID	Position Label	Camera Information	Image Resolution
Market1501 [21]	P	Real	32,668	6	1,501	-	2D	No	128×64
MSMT17 [16]	P	Real	126,441	15	4,101	-	2D	No	Vary
PersonX [12]	P	Synthetic	273,456	6	1,266	-	2D	No	Vary
RandPerson [15]	P	Synthetic	228,655	19	8,000	-	2D	No	Vary
UnRealPerson [20]	P	Synthetic	120,000	34	3,000	-	2D	No	Vary
DukeGroup [18]	G	Real	354	8	704	177	2D	No	Vary(max: 1392×630)
RoadGroup [18]	G	Real	324	2	1,099	162	2D	No	Vary (max: 450×255)
CSG [19]	G	Real	3,989	Vary	3,500	1,558	2D	No	Vary(max: 800×800)
City1M (Ours)	G	Synthetic	1,840,000	8	45,000	11,500	3D	Yes	1920×1080

called layout tokens, which can be modeled as follows.

$$\begin{cases} \mathcal{T}_x^\sigma = (t_x^0, t_x^1, \cdots, t_x^{\sigma-1}), \\ \mathcal{T}_y^\sigma = (t_y^0, t_y^1, \cdots, t_y^{\sigma-1}), \\ \mathcal{T}_d^\sigma = (t_d^0, t_d^1, \cdots, t_d^{\sigma-1}), \end{cases} \tag{4}$$

where the $t_m^n, m \in \{x, y, d\}, n \in \{0, 1, \cdots, \sigma-1\}$ is the 64-dim feature embedding.

Each token is initialized randomly at the beginning of training. With the update of network learning, tokens can represent the layout feature of the current location under the current dimension when the network training converges.

After spatial discretion, we can extract the 3D layout features from the layout relationship obtained in Sec. 3.1. For each member, we calculate the corresponding three tokens $\mathcal{I}_i^t$ of the i-th member as the layout feature according to positions obtained by Eq. (3), which can be described as follows.

$$\mathcal{I}_i^t = \left(\mathcal{T}_x^\sigma(\lfloor \frac{P_i(0)}{\sigma} \rfloor), \mathcal{T}_y^\sigma(\lfloor \frac{P_i(1)}{\sigma} \rfloor), \mathcal{T}_d^\sigma(\lfloor \frac{P_i(2)}{\sigma} \rfloor) \right). \tag{5}$$

After that, the member's appearance and layout features are concatenated and sent to the group transformer to obtain the group feature representation.

The proposed 3D token has the following three advantages. (1) 3D tokens consider the depth of members, which is ignored in previous methods. (2) 3D tokens discretize the X-Y-D space, which allows each token to represent a position within a certain neighborhood and is robust to possible layout changes or potential disturbances. (3) Our 3D token is very efficient, requiring only 3σ tokens. If some classic strategies are adopted, such as ViT [6], σ^3 tokens may be required. In this case, it is difficult to ensure that all tokens are adequately trained, resulting in poor layout features.

Finally, we need to provide supervision information for training the person and group transformers, including cross-entropy loss and hard triplet loss.

$$\mathcal{L}_c = -\frac{1}{M} \sum_{j=1}^{M} \sum_{i=1}^{C} y_{ji} \log(\widehat{y}_{ji}), \tag{6}$$

where M represents the member's number of the current batch, C represents the whole member classes, the indicator function y_{ji} equals to 1 when the j-th member belongs to the i-th class, and $\widehat{y}_{ji}$ is the prediction of the transformer.

$$\mathcal{L}_t = \frac{1}{M} \sum_{i=1}^{M} \left[d\left(f_i, f_i^+\right) - d\left(f_i, f_i^-\right) + m \right]_+, \tag{7}$$

where $d(\cdot, \cdot)$ represents the Euclidean distance between two features, $f_i/f_i^+/f_i^-$ represent the anchor/hard positive/hard negative feature in the current batch, $[\cdot]_+$ means $\max(\cdot, 0)$ and m is the margin.

$$\mathcal{L}_p = \mathcal{L}_c + \alpha \mathcal{L}_t, \tag{8}$$

$$\mathcal{L}_{all} = \mathcal{L}_g + \beta \mathcal{L}_p, \tag{9}$$

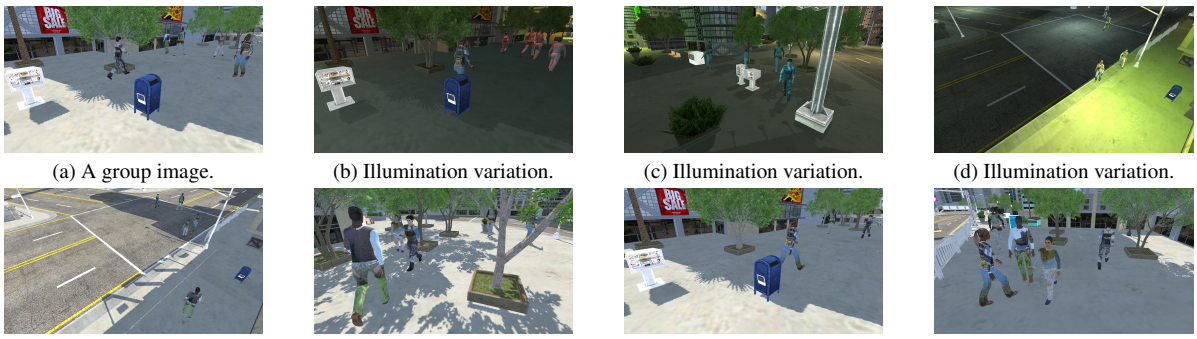
where the $\mathcal{L}_g$ is similar with the $\mathcal{L}_p$, α and β controls the balance between two different losses.

4. Synthetic Dataset: City1M

4.1. Human Production

We use MakeHuman [3] to generate diversified 3D person models, which is achieved by introducing randomization in both human body and clothing. First, we randomly assign the average human body in the attributes of age, weight, height, muscle, skin, hair and eyes to achieve the diversity of human bodies. Then, we randomly select 2,000 different images in the Google Landmarks dataset [17] to generate human upper and lower clothes.

We have produced a total of 45,000 3D human models. The above random settings of human body attributes and the random combinations of upper and lower clothes can fully guarantee the diversity of pedestrian appearance.



(a) A group image. (b) Illumination variation. (c) Illumination variation. (d) Illumination variation.

(e) Member scale variation. (f) Occlusion. (g) Number change of members. (h) Layout change of members.

Figure 3. Visualization of sampling diversity of the proposed City1M. The variations in Fig. 3b ∼ Fig. 3h are comparisons about Fig. 3a.

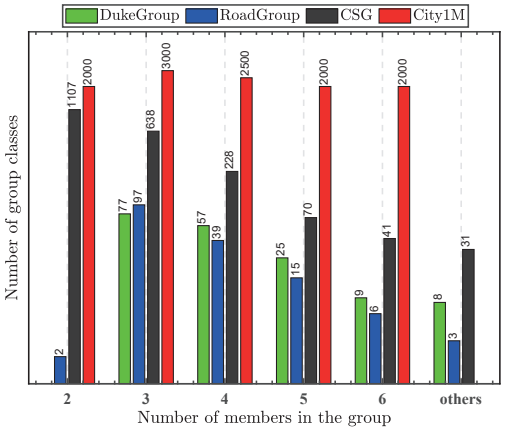


Figure 4. Statistics on the number of group classes of each number of members in the mainstream datasets. The value of the vertical axis has been shown by logarithm operation.

4.2. Dataset Construction

We implement the construction of City1M in Unity3D [13]. The motivation to propose City1M mainly consists of the following three aspects. First, the existing GReID datasets are very small. As shown in Tab. 1, the existing largest dataset CSG is only one-tenth of the traditional person ReID dataset Markets1501. In addition, the group images in CSG are not completely from the monitoring scenes, and some images are from the screenshots of the movies.

In contrast, all the images in the proposed City1M are captured from 8 street cameras in a modern city scene. Our City1M contains 1.84M images with a uniform resolution of 1920×1080, 45,000 persons and 11,000 groups, which are 470 times, 12 times and 7 times of CSG respectively. Besides, City1M simulates the potential variations that occur in real scenarios. As shown in Fig. 3, Fig. 3b ∼ Fig. 3d show the variations of illumination color caused by the change of day and night. Fig. 3e shows member scale variations caused by the different views. Fig. 3f mainly shows the inevitable occlusions of members in the monitoring scene, and also potentially shows variations of illumination intensity caused by the shadows. Fig. 3g and Fig. 3h show the number (Only 3 of 5 members are visible.) and layout variations (The member with blue trousers is on the right of Fig. 3a and the left of Fig. 3h) in GReID. The number variations in City1M also follow that images belonging to the same group class have more than 60% of the same members.

Second, the annotation of existing datasets is not abundant. Most datasets in Tab. 1 only provide the 2D plane coordinates of each person and do not provide additional information about the cameras, which is easy to obtain. The proposed City1M provides detailed 3D coordinates (The 2D coordinates of the imaging plane and the absolute depth information) for the position of each member. We also provide the position coordinates and shooting angle of each camera in the 3D scene, which is convenient for researchers to analyze the effect of camera networks. Furthermore, we also provide a time period label (captured in day or night) for each group image.

Finally, the effect of a synthetic dataset on GReID lacks exploration. As introduced in Sec. 2.2, synthetic datasets have been widely created to generate large-scale data in a low-cost way and can promote the performance of real datasets, which is less studied in GReID. City1M generates massive data with very low cost and simulates real monitoring scenarios and potential variations. Later experiments show that the model pretrained by City1M can further improve the performance in real scenes, which shows the effectiveness of the City1M.

5. Experiments

5.1. Datasets and Settings

Datasets. We evaluate the proposed 3DT on our City1M, DukeGroup [18], RoadGroup [18] and CSG [19]. The detailed information about the number of group images, cameras, person classes, and group classes have been shown in

Table 2. Performance comparisons with mainstream methods. Rank1, Rank5, Rank10 and mAP are reported (%). The "+" means the 3DT is pretrained on the proposed City1M.

Method	Publication	CSG				DukeGroup				RoadGroup			
		Rank1	Rank5	Rank10	mAP	Rank1	Rank5	Rank10	mAP	Rank1	Rank5	Rank10	mAP
CRRRO-BRO [22]	BMVC 2009	10.4	25.8	37.5	-	9.9	26.1	40.2	-	17.8	34.6	48.1	-
Covariance [4]	ICPR 2010	16.5	34.1	47.9	-	21.3	43.6	60.4	-	38.0	61.0	73.1	-
PREF [11]	ICCV 2017	19.2	36.4	51.8	-	30.6	55.3	67.0	-	43.0	68.7	77.9	-
BSC+CM [23]	ICIP 2016	24.6	38.5	55.1	-	23.1	44.3	56.4	-	58.6	80.6	87.4	-
LIMI [18]	MM 2018	-	-	-	-	47.4	68.1	77.3	-	72.3	90.6	94.1	-
DotGNN [8]	MM 2019	-	-	-	-	53.4	72.7	80.7	-	74.1	90.1	92.6	-
GCGNN [24]	TMM 2020	-	-	-	-	53.6	77.0	91.4	-	81.7	94.3	96.5	-
MGR [10]	TCYB 2021	57.8	71.6	76.5	-	48.4	75.2	89.9	-	80.2	93.8	96.3	-
MACG [19]	TPAMI 2020	63.2	75.4	79.7	-	57.4	79.0	90.3	-	84.5	95.0	96.9	-
DotSCN [9]	TCSVT 2021	-	-	-	-	86.4	98.8	98.8	-	84.0	95.1	96.3	-
3DT (Ours)	-	92.9	97.3	98.1	92.1	83.0	98.9	99.9	89.8	91.4	97.5	98.8	94.3
3DT+ (Ours)	-	**95.1**	**97.7**	**98.6**	**94.4**	**90.9**	**99.9**	**99.9**	**94.1**	**93.8**	**97.5**	**98.8**	**94.8**

Tab. 1. Similar to the protocol in [10], the training and testing set of DukeGroup and RoadGroup are equally split. Following the protocol in [19], 859/699 groups of 1,558 groups are split for training/testing. If there is no additional claim, we do not use any extra data when training on each dataset for fair comparison. The Cumulative Matching Characteristics (CMC) at Rank-1, Rank-5, Rank-10, and mean Average Precision (mAP) are used as evaluation metrics.

Settings. We adopt the standard ViT-Base [6], pretrained on ImageNet [5], as person transformer. For the group image, we crop all the members by the given bounding box and resize them to 256×128. In the training stage, we apply the random horizontal flip and random erasing. Each mini-batch is sampled with 16 group identities, and each group identity selects 4 images. We choose SGD [2] as the optimizer. The cosine annealing learning rate strategy is adopted. The initial learning rate is 2e-3, and the minimum learning rate is 1.6e-4. The weight decay is 1e-4. In the testing stage, we do not use any data augmentation and re-ranking. The Euclidean distance is applied. All ablation studies, parameter analyses, and visualizations have been conducted on the RoadGroup dataset.

5.2. Performance

We evaluate the proposed method against the existing methods on three GReID datasets. As shown in Table Tab. 2, the existing methods are divided into two groups: hand-crafted methods and deep learning methods. Note that DotSCN in deep learning method uses extra datasets for auxiliary training. We regard MACG as the best method for single dataset training and DotSCN as the best method for multiple dataset training. We also evaluate the proposed methods for single dataset and multiple dataset settings, called 3DT and 3DT+ respectively. 3DT+ first pretrains on City1M, and then finetunes and tests on each dataset.

Three conclusions can be drawn from Tab. 2. First, The proposed method in this paper achieves the state-of-the-art performances in single dataset training. Compared with MACG, 3DT exceeds 29.7%/35.6%/6.9% Rank1 on CSG/DukeGroup/RoadGroup datasets. Even without extra dataset, the 3DT surpasses DotCSN in most cases, which demonstrates the superiority of our method.

Compared with the existing methods, the advantages of 3DT/3DT+ mainly come from the following two aspects. (1) 3DT performs layout feature extraction based on the position in 3D space, rather than the 2D position of the imaging plane. Because the depth information is introduced into the 3D layout, the ambiguity of 2D layout in some scenes can be eliminated. (2) 3DT is a transformer-based framework. 3DT can model the layout features with the help of the layout tokens, which is difficult to extract in the traditional CNN-GNN framework.

Second, pretraining on the City1M dataset can further improve the performance (multiple dataset training). Compared with 3DT, the performance of 3DT+ on CSG, DukeGroup and RoadGroup dataset is further improved by 2.2%/2.3%, 7.9%/4.3% and 2.4%/0.5% on Rank1/mAP. This result demonstrates that the difference between the City1M and the real datasets is small enough. By pretraining on City1M, group prior knowledge can be transferred and model performance can be enhanced in real datasets.

Finally, the performance of 3DT/3DT+ can be improved on both large (CSG) and small (DukeGroup and RoadGroup) datasets, which shows that our method is robust to the scale of the dataset.

5.3. Ablation Study

As shown in Tab. 3, we analyze the each case of ignoring different spatial information, using only 1D position, 2D position and 3D position. Three conclusions can be drawn. First, if no layout features are considered, the performance is not satisfactory. In this case, the model distinguishes different group classes only by the appearance features, which will obtain high retrieval similarity between the hard negative samples with similar appearance.

Second, only using the 1D layout of X or Y can improve

Table 3. Ablation study of the 3D layout modeling. X, Y and D represent the three dimensions of the reconstructed 3D space. Rank1 and mAP are reported (%).

Type	X	Y	D	Rank1	mAP
None				88.89	91.25
1D	✓			88.89	91.71
		✓		88.89	92.03
			✓	87.65	91.49
2D	✓	✓		90.12	92.88
	✓		✓	88.89	91.36
		✓	✓	87.65	91.42
3D	✓	✓	✓	91.36	94.27

Table 4. Parameter analysis for sampling in 3D layout. Rank1 and mAP are reported (%).

σ	2	5	**10**	20	50
Rank1	90.12	91.36	**91.36**	91.36	90.12
mAP	92.92	92.90	**94.27**	93.34	92.62

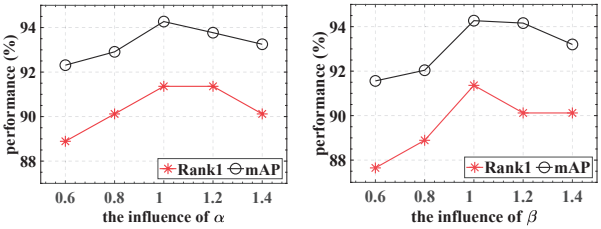
the performance. Specifically, the performance of 1D-X and 1D-Y is improved by 0.46% and 0.78% mAP because the positions of X and Y are true annotations and contain extra information that is different from the appearance. Introducing more prior knowledge will result in more performance gains. However, the performance benefit of using only D is not obvious. This phenomenon is because the information in the D dimension is obtained by estimation, which means that the D dimension itself is not completely accurate. Similar phenomenons can be found in 2D types. The performance improvement of using X-Y is greatest, X-D and Y-D will be limited by not using the whole prior knowledge but also introducing inaccurate estimation information.

Finally, the best performance is achieved when X,Y, and D are jointly adopted. Compared to the strategy that ignores layout modeling (Row1 in Tab. 3), using layout modeling brings an extra 2.47%/3.02% Rank1/mAP. Compared to strategies with 2D layout modeling (Row5 in Tab. 3), additional D information brings an extra 1.24%/1.39% Rank1 /mAP, which fully proves the superiority of our method.

5.4. Parameter Analysis

The influence of σ. Hyperparameter σ controls the discrete granularity of reconstructed 3D space in layout modeling. The larger σ corresponds to more fine-grained spatial discretization, which also means that more layout tokens need to be used to represent layout features. As shown in Tab. 4, when σ increases from 2 to 10, the performance also gradually increases and the best performance is achieved at $\sigma = 10$. This shows that small σ is rough for discretization so that members with a long distance use the same layout tokens, resulting in limited performance.

When σ is further increased to 50, the performance begins to decline, which shows that too large σ leads to a large number of tokens and the model cannot guarantee that each token is adequately trained.



(a) Hyperparameter α. (b) Hyperparameter β.

Figure 5. Parameter analysis for loss functions.

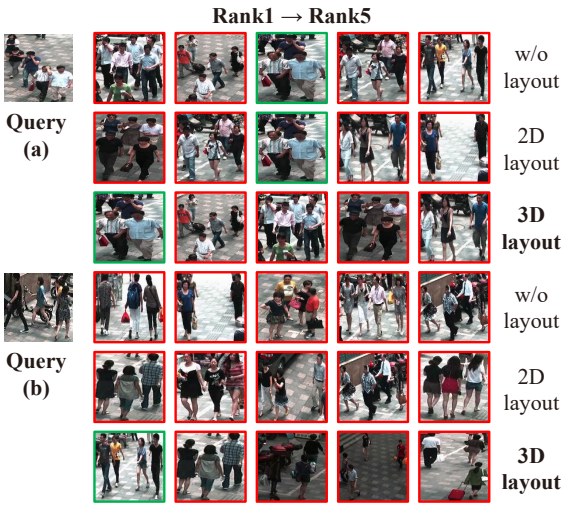


Figure 6. Visualization of top five retrieval results. Each row represents a different method, where "3D layout" is our method. Note that each query has only one correct image in the gallery. The green/red bounding box represents the correct/wrong matching.

The influence of α. Hyperparameter α controls the contribution of the cross-entropy loss and the triplet loss. As shown in Fig. 5a, our method achieves the best performance when $\alpha = 1.0$. Too large or small α will lead to performance degradation, which shows that the model tends to have the same contribution to these two losses.

The influence of β. Hyperparameter β controls the contribution of single person classification and group classification. As shown in Fig. 5b, our method achieves the best performance when $\beta = 1.0$, which is also intuitive because identifying individual members and groups is equally important for GReID.

5.5. Visualization

In Fig. 6, we enumerate two visual retrieval examples to prove the superiority of our method. In query (a), the correct gallery only has three members in common with

Table 5. The performance comparisons (%) of the different layout modeling strategies. #tokens represents the number of tokens required for each strategy.

strategy	#tokens	Rank1	mAP
Variation1	σ^3	88.89	92.28
Variation2	3	90.12	92.15
Ours	3σ	**91.36**	**94.27**

query, which cannot be addressed with only appearance. 2D-based method is also not ideal, because layout ambiguity will misjudge that the member with dark clothes in the correct gallery is very close to the member with light clothes. Our method combines appearance modeling with 3D layout modeling to obtain correct retrieval results.

In query (b), a group member in the query disappears, and another member is occluded by a passers-by. The method of "without layout" or "2D layout" can hardly deal with this situation. Our method can extract the appearance and layout features of the remaining two people, which leads to a high similarity matching.

5.6. Discussion

The analysis of alternative layout strategies. In addition to our strategy of modeling layout features, we also design the other two optional strategies. Variation1 adopts independent tokens for each small cube in the discrete 3D space, so σ^3 tokens are required for sampling rate σ. Variation2 only considers three basis vectors on three dimensions and uses the linear combination of three basis vectors to express the layout features of each small cube.

The comparisons are shown in Tab. 5, which proves that our strategy is better than these two variations. The number of tokens required by the Variation1 is very large, which will lead to insufficient training. The performance of Variation2 is still limited, indicating that the layout space is not consistent with linear space. Our strategy achieves a balance between the cost of tokens and the performance and achieves the best performance with relatively few tokens.

The effect of pretrained on City1M. We analyze the cross-dataset evaluation of our City1M on other datasets and the performances have been shown in Fig. 7. If the model pretrained on City1M is directly tested on other datasets, Rank1 has exceeded MACG on CSG and DukeGroup. This shows that City1M already contains more diverse groups, which is close to the distribution of the real dataset. The RoadGroup provides the cropped image instead of the original images. Therefore, the token pretrained on City1M cannot directly satisfy the layout of RoadGroup.

DotSCN uses the extra Market1501 dataset. Compared with the 3DT+, using the City1M will bring more performance improvement, which shows that the City1M is more suitable for GReID and can be widely used in the pretraining stage to obtain better performance.

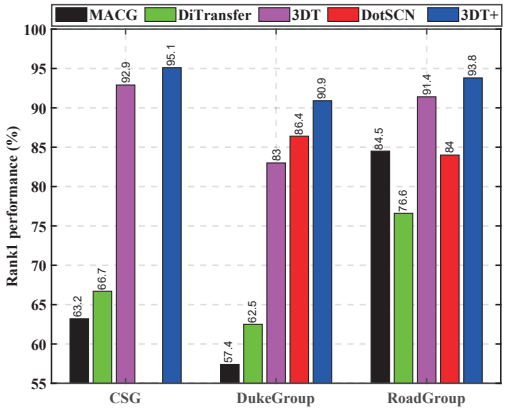


Figure 7. The effect of the proposed pretrained dataset City1M. Rank1 (%) is reported on three datasets. "DiTransfer" representatives directly test the model pretrained in City1M on each dataset.

Table 6. The performances of City1M (%).

Method	Setting	Rank1	Rank5	Rank10
3DT	Protocol@1	85.64	86.53	86.81
	Protocol@2	85.34	86.48	86.77
	Overall	85.49	86.51	86.79

The performance of City1M. We provide two evaluation protocols. Protocol@1 and protocol@2 split City1M equally into two parts. The training set and testing set in each protocol have 5.75K groups with 460K images respectively. Groups in the gallery that do not exist in query are regarded as distractions. Optionally, the last 10% of the training set can be treated as the validation set. Protocol@1 focuses on training in groups with fewer members and testing in groups with more members, and protocol@2 is just the opposite. The overall performances are the average of these two protocols, which are shown in Tab. 6.

6. Conclusion

In this paper, we extract group features with 3D layout modeling. Specifically, the proposed 3DT discretizes and samples the reconstructed 3D space. For each spatial cube, we use the combination of tokens with three dimensions as its layout feature. Furthermore, we propose a large-scale synthetic dataset, City1M, to alleviate the shortcomings of the existing GReID datasets. The experimental results show the superiority of our method and dataset.

Acknowledgments

This project was supported by the NSFC (62076258, 61902444), the Project of Natural Resources Department of Guangdong Province ([2021]34), and the Project of Ministry of Public Security of China (2019GABJC39).

References

[1] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *CVPR*, pages 4009–4018, June 2021. 3

[2] Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. 2012. 6

[3] Leyde Briceno and Gunther Paul. Makehuman: a review of the modelling framework. In *Congress of the International Ergonomics Association*, pages 224–232, 2018. 2, 4

[4] Yinghao Cai, Valtteri Takala, and Matti Pietikäinen. Matching groups of people by covariance descriptor. In *ICPR*, pages 2744–2747, 2010. 1, 2, 6

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 6

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3, 4, 6

[7] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *ICCV*, pages 15013–15022, 2021. 2

[8] 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 *ACM MM*, pages 1888–1896, 2019. 2, 6

[9] Ziling Huang, Zheng Wang, Chung-Chi Tsai, Shin'ichi Satoh, and Chia-Wen Lin. Dotscn: Group re-identification via domain-transferred single and couple representation learning. *IEEE TCSVT*, 31(7):2739–2750, 2021. 1, 2, 6

[10] 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 TCYB*, 51(3):1478–1492, 2021. 2, 6

[11] Giuseppe Lisanti, Niki Martinel, Alberto Del Bimbo, and Gian Luca Foresti. Group re-identification via unsupervised transfer of sparse features encoding. In *ICCV*, pages 2468–2477, 2017. 2, 6

[12] Xiaoxiao Sun and Liang Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *CVPR*, pages 608–617, 2019. 2, 4

[13] Unity Technologies. Unity3D: Cross-platform 3D engine, 2021. 2, 5

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, volume 30, 2017. 2

[15] Yanan Wang, Shengcai Liao, and Ling Shao. Surpassing real-world source training data: Random 3d characters for generalizable person re-identification. In Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann, editors, *ACM MM*, pages 3422–3430, 2020. 2, 4

[16] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer GAN to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018. 4

[17] Tobias Weyand, André Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2 - A large-scale benchmark for instance-level recognition and retrieval. In *CVPR*, pages 2572–2581, 2020. 4

[18] Hao Xiao, Weiyao Lin, Bin Sheng, Ke Lu, Junchi Yan, Jingdong Wang, Errui Ding, Yihao Zhang, and Hongkai Xiong. Group re-identification: Leveraging and integrating multi-grain information. In *ACM MM*, pages 192–200, 2018. 2, 4, 5, 6

[19] 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 TPAMI*, 2020. 2, 4, 5, 6

[20] Tianyu Zhang, Lingxi Xie, Longhui Wei, Zijie Zhuang, Yongfei Zhang, Bo Li, and Qi Tian. Unrealperson: An adaptive pipeline towards costless person re-identification. In *CVPR*, pages 11506–11515, 2021. 2, 4

[21] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 4

[22] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Associating groups of people. In *BMVC*, pages 1–11, 2009. 1, 2, 6

[23] Feng Zhu, Qi Chu, and Nenghai Yu. Consistent matching based on boosted salience channels for group re-identification. In *ICIP*, pages 4279–4283, 2016. 1, 2, 6

[24] Ji Zhu, Hua Yang, Weiyao Lin, Nian Liu, Jia Wang, and Wenjun Zhang. Group re-identification with group context graph neural networks. *IEEE TMM*, pages 1–1, 2020. 1, 2, 6