Cross-Resolution Person Re-Identification via Deep Group-Aware Representation Learning

Xiang Ye
College of Automation
Nanjing University of Posts and Telecommunications
Nanjing, China
Email: leafyewow@gmail.com

Guangwei Gao
Institute of Advanced Technology
Nanjing University of Posts and Telecommunications
Nanjing, China
Email: csggao@gmail.com

***Abstract*—Person re-identification (Re-ID) aims to identify the same person from samples taken by different cameras. However, in practical application scenarios, due to the quality of the camera equipment and the distance between the camera and the pedestrian, the captured pedestrian images usually have different resolutions, which will cause the mismatch problem of person Re-ID. To mitigate the resolution discrepancy issue, in this paper, we propose a method called deep group-aware representation learning (DGRL) for effective Re-ID. Firstly, We use the residual Transformer block in the feature extraction stage to thoroughly extract richer local and global information from variable resolution shallow images. Then our proposed multi-layered group-aware representation (MGAR) scheme can generate diverse representations different from the main branch, thereby improving the representation capability of the deeply embedded features. In addition, we calculate the kullback leibler divergence loss (KLDivLoss) values on the probability prediction outputs of any two branches, forcing the entire network to be well optimized. Plenty of evaluations on four benchmark datasets have demonstrated the effectiveness of our method.**

I. Introduction

Person re-identification (Re-ID) tasks aim to identify the same person from images captured by different cameras, which is a potential research area in the field of pattern recognition community [1]–[3]. Due to the complexity of the surveillance video scene, the main challenge for person Re-ID comes from the large variants of persons, such as posture, occlusion, clothes, background clutters, etc. The development of deep convolutional neural networks has introduced more powerful and robust representations for pedestrian image recognition, increasing the performance of Re-ID algorithms to a new level. Recently, some deep Re-ID methods have achieved satisfactory matching accuracy [4]–[7].

However, in real-world scenarios, due to the quality problems of camera equipment and the distance between the camera and pedestrians, the pedestrian images captured by the camera usually have poor quality, such as the low-resolution (LR) issue. The images provided by some equipment (usually viewed as the retrieval gallery) may be high-resolution (HR), leading to another important research topic, i.e., cross-resolution person re-identification. Fig. 1 shows the differences between general and cross-resolution person Re-ID.

In past few years, many cross-resolution person Re-ID algorithms have been presented to tackle the above problems.

Fig. 1: Illustration of person Re-ID problem: (a) general person Re-ID task in the ideal scenarios, (b) cross-resolution person Re-ID task in cross-camera scenarios.

Early work mainly used metric learning scheme to extract the common feature representations from both LR and HR images to realize image matching [8], [9]. However, due to the loss of fine-grained detail information in LR person images, the performance of these methods is limited. Then, many researchers try to introduce the image super-resolution (SR) methods into cross-resolution effective person Re-ID tasks. SING [10] firstly explored SRCNN [11] method as the resolution restoration scheme. Subsequently, various advanced SR networks [12]–[14] were introduced to further optimize the framework. These models usually adopt a joint training scheme that cascades the SR module and the Re-ID sub-network. However, this design is subject to ineffective model training since it is significantly more difficult to back-propagate the gradient through such a cascaded complex model [15]. Therefore, the compatibility features between the SR model and the Re-ID model may be poor. Recently, several novel and effective methods have been proposed [15], [16], which effectively improve the matching accuracy and performance, but it is still far from practical applications.

Since the SR methods are introduced into the cross-resolution person Re-ID tasks, it seems unreasonable to pay too much attention to the joint training of the cascaded SR scheme and Re-ID sub-network. In addition, in some cases, complex SR networks are not necessarily better than simple SR ones in the cross-resolution person Re-ID tasks. We need to put more thought into the deep semantic information and feature information extraction of images [16].

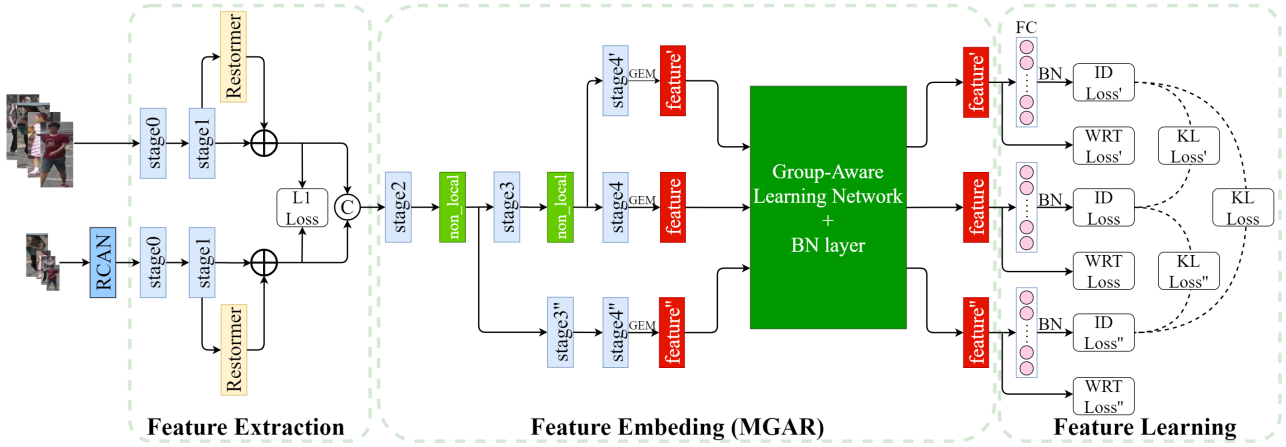

Fig. 2: The framework of our deep group-aware representation learning (DGRL) for cross-resolution person Re-ID tasks.

By considering the above analysises and observations, we design a deep group-aware representation learning (DGRL) model for cross-resolution person Re-ID tasks. We introduce a residual Transformer block and multi-layered group-aware representation (MGAR) scheme to drastically enhance the discriminant ability of the deeply embedded features.

The primary contributions of our method is listed as follows:

- We use the residual Transformer block in the feature extraction stage to extract richer local and global information from the shallow images. As we know, this is the first attempt to apply the Transformer to the effective cross-resolution person Re-ID tasks.
- Our multi-layered group-aware representation (MGAR) scheme can generate abundant representations that are different from the main branch, thereby enhancing the quality of the deeply embedding features.
- In addition, we design and calculate the KLDivLoss on the probability prediction outputs of any two branches, forcing the entire network to be well optimized. Extensive experiments have revealed the superiority of our proposed DGRL method.

II. Related Works

In this section, we briefly introduce some works related to our work, that is general person Re-ID and cross-resolution person Re-ID.

A. Person Re-ID

The goal of person Re-ID is to solve the pedestrian matching problem between non-overlapping discrepant cameras [17]–[19]. With the development of deep convolutional neural networks, many deep Re-ID solutions have been presented to deal with challenges such as posture, occlusion, clothing, background clutters, etc. To solve the problem of pedestrian posture change and misalignment, Wei et al. [20] used four key points to divide human body into three parts: head, upper body, and lower body, and applied descriptor learning to learn local information and global information of the three parts respectively. To handle the occlusion problem, Wang et al. [21] proposed the adaptive-direction graph convolutional layer to enhance the semantic information transmission and the cross-graph alignment layer to learn the correspondence between nodes. Yang et al. [22] designed a method to solve the problem of clothing change by taking the center of the pedestrian outline sketch as the pole and scanning all the outline information of the whole person by constantly changing the angle and radius. Kalayeh et al. [23] used the branch of human semantic passing to weigh the features of the backbone to separate the foreground information from the background information.

B. Cross-Resolution Person Re-ID

To solve the resolution mismatch issue, some cross-resolution person Re-ID approaches have been designed in recent years. Jiao et al. [10] introduced SRCNN into the Re-ID network for the first time, and designed a hybrid depth convolution neural network to improve the matching performance. Chen et al. [24] used the generative adversarial network (GAN) [25] to extract the resolution invariant representations for Re-ID task. Han et al. [26] proposed a model to forecast the scale factor according to the image content and adaptively recover the missing details of the image to solve the issue of resolution mismatch. Chen et al. [15] explored the potential relationship between SR and Re-ID module and applied it as a constraint to enhance the compatibility of cascaded networks. Recently, the PS-HRNET method [16] designed the VDSR-CA module and a new presentation head, which used a pseudo-siamese architecture to narrow the discrepancy of feature distributions between the LR and HR person images.

III. Proposed DGRL

In this part, we first introduce the overall framework of the model, then introduce the implementation details of each module, and the loss functions in the framework.

A. Framework Overview

For the input images, we let h and l represent the HR and LR images. Thus, a set of N HR images is denoted as

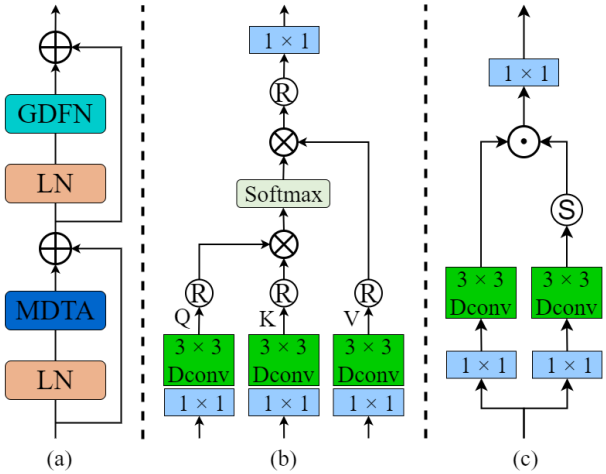

Fig. 3: (a) The architecture of the Restormer block. (b) Structure of MDTA. (c) Structure of GDFN.

$\mathcal{D}_h = \{x_h^i, y^i\}_{i=1}^N$, where $x_h^i \in \mathbb{R}^{C\times H\times W}$ and $y^i \in \mathbb{R}$ are the i^{th} HR image and its label, respectively. The C, H, W denote channel, height and width respectively. We conduct random down-sampling for each HR sample, and the sampling factor $r \in \{2, 3, 4\}$. The created corresponding LR samples are represented as $\mathcal{D}_l = \{x_l^i, y^i\}_{i=1}^N$.

As depicted in Fig. 2, our proposed DGRL method uses Resnet50 [27] as the backbone network. Firstly, before the feature extraction module, we feed the LR images into the pretrained RCAN [28] network to generate corresponding SR ones, rather than training the super-resolution network as a part of the overall framework. Secondly, HR and SR images are fed to the dual-branch feature extraction module without shared parameters, and shallow image features are extracted. Then the features extracted from the double branches are sent into the feature embedding module with shared parameters. Finally, we use some loss functions to guide the efficient feature learning process. These modules will be described in detail in the following sections.

B. Super-Resolution and Feature Extraction Module

1) Super-Resolution Module: PS-HRNet [16] introduced the channel attention mechanism into VDSR [29] and named it VDSR-CA as the SR module. This joint training strategy of cascading SR and Re-ID model will make the gradient back-propagation of the model more difficult, thus affecting the training effect [15]. Therefore, to avoid the above problem, before the feature extraction module, we send the LR images into the pretrained RCAN model to obtain the SR images. The generated corresponding SR samples are represented as $\mathcal{D}_s = \{x_s^i, y^i\}_{i=1}^N$.

2) Feature Extraction Module: The feature extraction module mainly consists of two branches that do not share the parameters. We send the paired HR image x_h^i and SR image x_s^i into the feature extraction module for shallow feature extraction. To extract richer local and global information from shallow images, we introduce Restormer block [30] and embed it into a residual structure. As we know, this is the first attempt that Transformer is introduced into cross-resolution person Re-ID tasks.

As depicted in Fig. 3 (a), the Restormer block is mainly composed of the Multi-Dconv Head Transposed Attention (MDTA), Gated-Dconv Feed-Forward Network (GDFN), and two LayerNorm (LN) layers. The LN layer is introduced before both MDTA and GDFN modules and a residual connection is applied after each module. The architecture of MDTA is shown in Fig. 3 (b), which uses 1×1 convolutions to assemble pixel-wise cross-channel context and 3×3 depth-wise convolutions to distill channel-wise spatial context. Given a normalized tensor $X \in \mathbb{R}^{H\times W\times C}$, the *query*, *key* and *value* matrices Q, K and V are formulated as follows:

$$Q = W_p^Q W_d^Q X, \quad K = W_p^K W_d^K X, \quad V = W_p^V W_d^V X, \tag{1}$$

where $W_p^{(\cdot)}$ denotes the 1×1 point-wise convolution and $W_d^{(\cdot)}$ represents the 3×3 depth-wise convolution. Next, we reshape the query and key projections such that their dot-product interaction generates a transposed attention map. Therefore, the MDTA process is calculated as follows:

$$\begin{aligned} &\text{MDTA}(X) = W_p \text{ Attention } (\hat{Q}, \hat{K}, \hat{V}), \\ &\text{Attention}(\hat{Q}, \hat{K}, \hat{V}) = \hat{V} \cdot \text{Softmax}(\hat{K} \cdot \hat{Q}/\alpha), \end{aligned} \tag{2}$$

where matrices $\hat{Q} \in \mathbb{R}^{HW\times C}$, $\hat{K} \in \mathbb{R}^{C\times HW}$, and $\hat{V} \in \mathbb{R}^{HW\times C}$ are acquired after the tensor recombination operations. Here, α denotes a learnable scaling variable to restrain the importance of the dot product of $\hat{Q}$ and $\hat{K}$.

The structure of GDFN is depicted in Fig. 3 (c). The gating scheme is designed as the element-wise product of two parallel paths of the linear Transformation layers, one of which is activated by the Gaussian Error Linear Units (GELU) non-linearity [31] operation. Given an acquired tensor $X \in \mathbb{R}^{H\times W\times C}$, GDFN is calculated as

$$\begin{aligned} &\text{GDFN}(X) = W_p^0 \text{ Gating } (X), \\ &\text{Gating}(X) = \phi\left(W_d^1 W_p^1 X\right) \odot W_d^2 W_p^2 X, \end{aligned} \tag{3}$$

where $\odot$ denotes the element-wise multiplication, ϕ represents the GELU non-linearity. The overall process is summarized as

$$\begin{aligned} X &= \text{MDTA}(\text{LN}(X)) + X, \\ X &= \text{GDFN}(\text{LN}(X)) + X. \end{aligned} \tag{4}$$

At last, the final representations f_h^i, f_s^i in feature extraction module is obtained by adding the input and output of the Restormer block. The simple L_1 loss is utilized to constraint the feature extraction procedure, which is defined as

$$L_1 = \sum_{i\in\mathcal{B}} \begin{cases} 0.5\left(f_h^i - f_s^i\right)^2, & \left|f_h^i - f_s^i\right| < 1, \\ \left|f_h^i - f_s^i\right| - 0.5, & \text{otherwise}. \end{cases} \tag{5}$$

C. Feature Embedding Module

As illustrated in Fig. 2, we propose a novel multi-layered group-aware representation (MGAR) scheme in the feature embedding module. It is a huge challenge to recognize

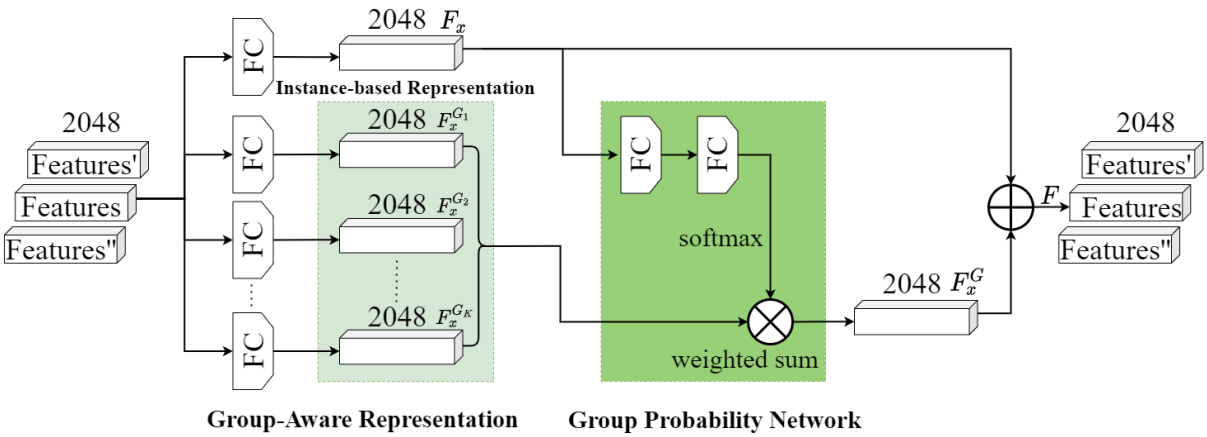

Fig. 4: The architecture of group-aware learning network (GALN).

thousands of personal images with different resolutions and perspectives. Using group-aware feature representations encourages models to learn more discriminative representations from various person features.

Firstly, the double branch features from the feature extraction module are spliced and sent to the feature embedding module. Partly inspired by the DHM [32], to learn a strong discriminative embedding space, we use complex auxiliary branches to improve the representation abilities of the feature embedding. We derive well-designed side branches from the Resnet50 middle tier. Each branch appears differently from that in Resnet50, which not only retains the representation of the backbone, but also generates more diverse representations along its own path. Specifically, we generate a side branch from stage2 and stage3 respectively, and each side branch is made up of building blocks (residual blocks in ResNet50) of the same type. Furthermore, both sides and the main branch keep the same down-sampling rate through their own pathway to the ending group-aware attention module. We use generalized-mean (GEM) pooling [33], which is a learnable layer, to capture domain-specific discriminative features.

Then, the obtained multi-layer features are fed into the group-aware learning network (GALN) (as illustrated in Fig. 4). GALN computes and uses both instance-based and group-aware representations. The instance-based representation F_x is obtained by the fully connected (FC) layer, and the K group-aware representations $F_x^{G_K}$ can also be obtained by deploying parallel FC layers. In addition, the group probabilities are calculated from the instance-based representation vector by deploying a group probability network that consists of two FC layers and a softmax layer. Finally, group probability determines which representation is more important to the final coding feature, and the assemble of multiple group-aware representation F_x^G and the instance-based representation F_x are added to obtain the final representation F. Similarly, we can get other two representations $F^{'}$ and $F^{"}$.

Our MGAR not only enriches image feature representation, but also improves the quality of the features through group-aware strategy, greatly promoting the performance of cross-resolution person Re-ID.

D. Feature Learning Module

1) Identification (ID) Loss: To improve the feasibility and effectiveness of the classification, we apply the cross-entropy loss as the ID loss. Given an input image, we indicate y as the true ID label and p_i as the ID prediction logits of the i_{th} class. For ID loss, label smoothing technology is adopted to prevent model over-fitting. We calculate the ID loss of the three branches respectively. Each ID loss is calculated as follows:

$$\begin{aligned} L_{id} &= \sum_{i=1}^{N} -q_i \log\left(p_i\right), \\ \text{s.t. } q_i &= \begin{cases} 1-\frac{N-1}{N}\xi, & y=i, \\ \frac{\xi}{N}, & y\neq i, \end{cases} \end{aligned} \tag{6}$$

where N denotes the total number of identities in the training set, and ξ is a soft-margin to reduce the model over-confidence. In this work, ξ is set to 0.1. We compute all ID losses by

$$L_{ID} = L_{id} + \alpha_1 L_{id'} + \beta_1 L_{id"}, \tag{7}$$

where α_1 and β_1 are two variables to balance the importance of $L_{id'}$ and $L_{id"}$, respectively.

2) Weighted Regularization Triplet (WRT) Loss: Except for the above essential cross-entropy loss with label smoothing, we also integrate WRT [38] loss to improve the model's performance on hard samples. We also calculate WRT loss of the three branches respectively. Each WRT loss is calculated as follows:

$$\begin{aligned} L_{wrt}(i,j,k) &= \log\left(1+\exp\left(w_i^p d_{ij}^p - w_i^n d_{ik}^n\right)\right), \\ w_i^p = \frac{\exp\left(d_{ij}^p\right)}{\sum_{d^p\in\mathcal{P}}\exp\left(d^p\right)}, & w_i^n = \frac{\exp\left(-d_{ik}^n\right)}{\sum_{d^n\in\mathcal{N}}\exp\left(-d^n\right)}, \end{aligned} \tag{8}$$

where (i,j,k) denotes the mined hard triplet within each training batch. $\mathcal{P}$ represents the positive sample set corresponding to the anchor point, and $\mathcal{N}$ represents the negative sample

TABLE I: Experiments (%) for cross-resolution person Re-ID tasks. The top two results are highlighted and underlined.

Module	MLR-Market-1501			MLR-CUHK03			MLR-VIPeR			CAVIAR		
	Rank1	Rank5	Rank10	Rank1	Rank5	Rank10	Rank1	Rank5	Rank10	Rank1	Rank5	Rank10
CamStyle [34]	74.5	88.6	93	69.1	89.6	93.9	34.4	56.8	66.6	32.1	72.3	85.9
PyrNet [35]	83.8	93.3	95.6	83.9	97.1	98.5	-	-	-	43.6	79.2	90.4
FD-GAN [36]	79.6	91.6	93.5	73.4	93.8	97.9	39.1	62.1	72.5	33.5	71.4	86.5
SING [10]	74.4	87.8	91.6	67.7	90.7	94.7	33.5	57	66.5	33.5	72.7	89
CAD-Net [37]	83.7	92.7	95.8	82.1	97.4	98.8	43.1	68.2	77.5	42.8	76.2	91.5
INTACT [15]	88.1	95	96.9	86.4	97.4	98.5	46.2	73.1	81.6	44	81.8	93.9
PRI [26]	84.9	93.5	96.1	85.2	97.5	98.8	-	-	-	43.2	78.5	91.9
PCB+PRI [26]	88.1	94.2	96.5	86.2	97.9	99.1	-	-	-	44.3	83.7	94.8
PyrNet+PRI [26]	86.9	93.8	96.4	86.5	97.7	99.1	-	-	-	45.2	84.1	94.6
PS-HRNet [16]	91.5	96.7	97.9	92.6	98.3	99.4	**48.7**	73.4	81.7	48.2	**84.5**	**96.3**
DGRL (Ours)	**93.9**	**98.0**	**98.8**	**98.1**	**99.6**	**99.6**	45.9	**73.7**	**83.9**	**52.4**	78.0	84.8

set corresponding to the anchor point. d_{ij}^{p} and d_{ij}^{n} represents the similarity from the anchor point to the farthest positive and nearest negative sample, respectively. The aforementioned weighted regularization not only inherits the merits of relative distance optimization between the negative and positive pairs, but also avoids importing any additional margin variables. WRT loss has no parameters at all, which makes it more flexible and adaptive. We compute all WRT losses by

$$L_{WRT} = L_{wrt} + \alpha_2 L_{wrt'} + \beta_2 L_{wrt"}, \tag{9}$$

where α_2 and β_2 are two variables to balance the importance of $L_{wrt'}$ and $L_{wrt"}$, respectively.

3) KLDivLoss: To prevent the features generated by the side branches from being inconsistent with the main branch, and also to increase the opportunity of knowledge sharing, we calculate KLDivLoss on the probability prediction outputs of any two branches, forcing the entire network to be well optimized. The probability of class n for a sample x_i given by the main branch is calculated as

$$p_1^n(x_i) = \frac{\exp(s_1^n)}{\sum_{n=1}^{N} \exp(s_1^n)}, \tag{10}$$

where s_1^n is the output of the "classfier" layer in model. Similarly, we can get three predictions p_1, p_2 and p_3. The KLDivLoss of p_1, p_2 is computed as

$$D_{KL}(p_2 \| p_1) = \sum_{i=1}^{M} \sum_{n=1}^{N} p_2^n(x_i) \log \frac{p_2^n(x_i)}{p_1^n(x_i)}. \tag{11}$$

We compute all KLDivLosses by

$$D_{KL} = \sum_{j \neq i}^{3} \sum_{i=1}^{3} D_{KL}(p_j \| p_i). \tag{12}$$

4) Overall loss: The holistic loss function L for optimizing the presented DGRL is given by

$$L = \gamma L_1 + L_{ID} + L_{WRT} + D_{KL}, \tag{13}$$

where γ represents a parameter to balance the importance of L_1 in all loss functions.

IV. Experiments and Discussions

A. Datasets Statements

We assess the effectiveness of our DGRL on three synthetic and one real cross-resolution Re-ID datasets. Three synthetic cross-resolution datasets, which contain multiple low resolutions (MLR), are constructed from the common person Re-ID datasets: Market-1501 [39], CUHK03 [40], and VIPeR [41]. The four datasets are described as follows:

1) MLR-Market-1501: The widely used MLR-Market-1501 set contains 32,668 images from 1,501 persons acquired in 6 cameras. Inspired by [37], we pretreat samples from one camera using the identical down-sampling factor, while other image resolutions remain unchanged. We apply the widely used 751/750 training/test identity split criterion.

2) MLR-CUHK03: The MLR-CUHK03 set is composed of 14,096 images of 1,467 individuals taken by 10 (5 pairs) different cameras. Following [10], for each pair of cameras, the down-sampling rate $r \in \{2, 3, 4\}$ is randomly selected to down-sample the images captured by one camera, while the size of images captured by other cameras remains unchanged. We apply the widely adopted 1,367/100 training/testing identity split criterion.

3) MLR-VIPeR: The MLR-VIPeR dataset includes 1264 images of 632 persons from 2 cameras. Following [10], we randomly down-sample all samples from one camera with the down-sampling rate $r \in \{2, 3, 4\}$, while the size of samples from other cameras remains unchanged. We adopt the widely used 316/316 training/testing identity split criterion.

4) CAVIAR: The CAVIAR [42] dataset contains 1,220 images of 72 individuals from 2 cameras. Following [10], we discard 22 person who are only showed up in one camera and use the 25/25 training/testing identity split. Unlike other synthetic datasets, CAVIAR dataset involves a variety of real cross-resolution images.

B. Experimental Settings

We evaluate the performance of our DGRL applying the cross-resolution person Re-ID settings [10], where the gallery

set contains HR samples while the probe (query) set contains LR samples. We employ the normal single-shot person Re-ID settings and adopt the average cumulative match characteristic (CMC) to assess the performance. We give the performance comparisons of ranks 1, 5, and 10. When testing, we only use the main branch features for matching and evaluation.

C. Implementation Details

We employ the PyTorch framework to implement our model. Before model training, LR images are obtained by randomly down-sampling HR images with the sampling factor $r \in \{2, 3, 4\}$ by the way of resampling using pixel area relation. We pretrain RCAN on the dataset and get the corresponding SR images. For the CAVIAR dataset, we directly send the images into the model. In the feature embedding module, we only use one Restormer block and set the number of multi-head attention to 4. As for the MGRN module, we generate a new branch from stage2 and stage3 respectively, and use 32 group-aware representations in the GALN.

Before training, all the HR and LR images are reshaped to have the size of $256 \times 128 \times 3$ for both model training and deployment. A mini-batch contains 48 pairs of images of 16 individuals, and each individual has 3 pairs of LR and HR images. We use SGD as the optimizer with the weight decay is 5×10^{-4} and the momentum is 0.9. We train the model 70 epochs in total. The learning rate is empirically set to 1.3×10^{-2}, which is decreased by one-tenth every 30 epochs. To reduce the impact of side branches, the hyper-parameters α_1, β_1, α_2 and β_2 are set to 0.5. The hyper-parameter γ is set to 5. We learn our network on a single NVIDIA GeForce RTX 3090 GPU with 24GB memory.

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

We compare our DGRL with plenty of competitive methods, which can be roughly categorized into two classes: (1) Traditional person Re-ID methods: CamStyle [34], PyrNet [35], and FD-GAN [36]; (2) Specially presented cross-resolution person Re-ID methods: SING [10], CAD-Net [37], INTACT [15], PRI [26], PCB+PRI [26], PyrNet+PRI [26], and PS-HRNet [16]. The comparison results are shown in TABLE I.

Quantitative results have verified that our DGRL obtains state-of-the-art performance on four MLR datasets. Specially, in contrast to the traditional person Re-ID methods, our DGRL improves 19.4%, 10.1% and 14.3% in terms of Rank-1 on MLR-Market1501 dataset, which shows that the ordinary person Re-ID methods can not effectively extract the useful information in LR images. Compared with the methods specially designed for cross-resolution person Re-ID problems, our proposed method shows superior performance over other competitive approaches, which explicitly reveals the importance of our deep group-aware representation learning scheme.

E. Ablation Studies

1) Effect of Loss Functions: We validate the effectiveness of all loss functions in this part. TABLE II reports the ablation results on the MLR-Market-1501 dataset. We can see that ID loss and triplet loss are very common in person Re-ID problems. After adding the L_1 loss, the distances between HR images and LR images can be further narrowed during the feature extraction module. KLDiv loss can force the entire network to be optimized in one direction.

TABLE II: Evaluation (%) of loss functions of our DGRL on the widely used MLR-Market-1501 dataset.

Supervision	Rank1	Rank5	Rank10
ID	89.9	96.3	97.8
ID + WRT	92.9	97.3	98.6
ID + WRT + L1	93.2	97.9	98.7
ID + WRT + L1 + KLDiv	**93.9**	**98.0**	**98.8**

TABLE III: Components analysis (%) of our DGRL on the MLR-Market-1501 dataset.

Non-local Attention	GALN	Restormer	Rank1	Rank5	Rank10
×	×	×	87.0	94.7	95.8
✓	×	×	87.5	95.0	96.8
✓	✓	×	93.3	98.0	98.7
✓	✓	✓	**93.9**	**98.0**	**98.8**

2) Component Analysis: We analyze the effectiveness of each component next. The first row of Table III shows a simple baseline. The second row in the table tabulates the effectiveness of adding non-local attention to the feature embedding module. Comparing the second and third row in TABLE III, we can observe that our GALN greatly enhances the quality of deep feature representations. From the last row in Table III we can discover that the Restormer block brings advantages to the network. The combination of all components significantly yields the best results.

V. Conclusions

In this paper, we proposed a new method called deep group-aware representation learning (DGRL) for cross-resolution efficient person Re-ID problem. In feature extraction module, we use the residual Transformer block to extract richer global information from shallow images. In feature embedding module, our multi-layered group-aware representation (MGAR) scheme can generate diverse representations that are different from the main branch, thereby improving the representation capability of deep embedding features. In addition, we calculate KLDivLoss on the probability prediction outputs of any two branches, forcing the entire network to be well optimized. Experimental results have shown that our model is superior to many existing cross-resolution person Re-ID methods.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant no. 61972212, the Natural Science Foundation of Jiangsu Province under Grant no. BK20190089. *(Corresponding author: Guangwei Gao)*

REFERENCES

[1] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.

[2] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2109–2118.

[3] Y. Huang, S. Zhang, H. Hu, D. Chen, and T. Su, "Resetting-label network based on fast group loss for person re-identification," *IEEE Access*, vol. 7, pp. 119 486–119 496, 2019.

[4] H. Ye, H. Liu, F. Meng, and X. Li, "Bi-directional exponential angular triplet loss for rgb-infrared person re-identification," *IEEE Transactions on Image Processing*, vol. 30, pp. 1583–1595, 2021.

[5] C. Zhao, X. Wang, W. Zuo, F. Shen, L. Shao, and D. Miao, "Similarity learning with joint transfer constraints for person re-identification," *Pattern Recognition*, vol. 97, p. 107014, 2020.

[6] G. Gao, H. Shao, F. Wu, M. Yang, and Y. Yu, "Leaning compact and representative features for cross-modality person re-identification," *World Wide Web*, pp. 1–18, 2022.

[7] G. Gao, Y. Yu, J. Yang, G.-J. Qi, and M. Yang, "Hierarchical deep cnn feature set-based representation learning for robust cross-resolution face recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2550–2560, 2022.

[8] G. Zhang, J. Yang, Y. Zheng, Z. Luo, and J. Zhang, "Optimal discriminative feature and dictionary learning for image set classification," *Information Sciences*, vol. 547, pp. 498–513, 2021.

[9] G. Zhang, H. Sun, F. Porikli, Y. Liu, and Q. Sun, "Optimal couple projections for domain adaptive sparse representation-based classification," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5922–5935, 2017.

[10] J. Jiao, W.-S. Zheng, A. Wu, X. Zhu, and S. Gong, "Deep low-resolution person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[11] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proceedings of the European Conference on Computer Vision*. Springer, 2014, pp. 184–199.

[12] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4681–4690.

[13] G. Gao, W. Li, J. Li, F. Wu, H. Lu, and Y. Yu, "Feature distillation interaction weighting network for lightweight image super-resolution," *arXiv preprint arXiv:2112.08655*, 2021.

[14] G. Gao, Z. Wang, J. Li, W. Li, Y. Yu, and T. Zeng, "Lightweight bimodal network for single-image super-resolution via symmetric cnn and recursive transformer," *arXiv preprint arXiv:2204.13286*, 2022.

[15] Z. Cheng, Q. Dong, S. Gong, and X. Zhu, "Inter-task association critic for cross-resolution person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2605–2615.

[16] G. Zhang, Y. Ge, Z. Dong, H. Wang, Y. Zheng, and S. Chen, "Deep high-resolution representation learning for cross-resolution person re-identification," *arXiv preprint arXiv:2105.11722*, 2021.

[17] J. Dietlmeier, J. Antony, K. McGuinness, and N. E. O'Connor, "How important are faces for person re-identification?" in *Proceedings of the International Conference on Pattern Recognition*. IEEE, 2021, pp. 6912–6919.

[18] M. Ye, H. Li, B. Du, J. Shen, L. Shao, and S. C. Hoi, "Collaborative refining for person re-identification with label noise," *IEEE Transactions on Image Processing*, vol. 31, pp. 379–391, 2022.

[19] J. Li, S. Zhang, Q. Tian, M. Wang, and W. Gao, "Pose-guided representation learning for person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 2, pp. 622–635, 2022.

[20] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "Glad: Global-local-alignment descriptor for pedestrian retrieval," in *Proceedings of the ACM international conference on Multimedia*, 2017, pp. 420–428.

[21] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun, "High-order information matters: Learning relation and topology for occluded person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6449–6458.

[22] Q. Yang, A. Wu, and W.-S. Zheng, "Person re-identification by contour sketch under moderate clothing change," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[23] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1062–1071.

[24] Y.-C. Chen, Y.-J. Li, X. Du, and Y.-C. F. Wang, "Learning resolution-invariant deep representations for person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8215–8222.

[25] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," *arXiv preprint arXiv:1406.2661*, p. arXiv:1406.2661, Jun. 2014.

[26] K. Han, Y. Huang, Z. Chen, L. Wang, and T. Tan, "Prediction and recovery for adaptive low-resolution person re-identification," in *Proceedings of the European Conference on Computer Vision*. Springer, 2020, pp. 193–209.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[28] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 286–301.

[29] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1646–1654.

[30] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," *arXiv preprint arXiv:2111.09881*, 2021.

[31] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[32] D. Li and Q. Chen, "Dynamic hierarchical mimicking towards consistent optimization objectives," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7642–7651.

[33] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018.

[34] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5157–5166.

[35] N. Martinel, G. Luca Foresti, and C. Micheloni, "Aggregating deep pyramidal representations for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 1–11.

[36] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, and H. Li, "Fd-gan: Pose-guided feature distilling gan for robust person re-identification," *arXiv preprint arXiv:1810.02936*, 2018.

[37] Y.-J. Li, Y.-C. Chen, Y.-Y. Lin, X. Du, and Y.-C. F. Wang, "Recover and identify: A generative dual model for cross-resolution person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8090–8099.

[38] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[39] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1116–1124.

[40] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proceedings of the European Conference on Computer Vision*. Springer, 2008, pp. 262–275.

[41] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 152–159.

[42] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3754–3762.