IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received October 13, 2020, accepted November 6, 2020, date of publication November 9, 2020, date of current version November 19, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3036985

A Discriminative Person Re-Identification Model With Global-Local Attention and Adaptive Weighted Rank List Loss

YONGCHANG GONG[1], LIEJUN WANG[1,2], YONGMING LI[1], AND ANYU DU[1]
[1]College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China
[2]Key Laboratory of Signal Detection and Processing, Xinjiang Uygur Autonomous Region, Xinjiang University, Xinjiang 830046, China

Corresponding author: Liejun Wang (wljxju@xju.edu.cn)

This work was supported in part by the Xinjiang Uygur Autonomous Region Natural Science Foundation Project under Grant 2020D01C058, in part by the National Natural Science Foundation Project under Grant 61771416 and Grant U1903213, in part by the Xinjiang Uygur Autonomous Region Graduate Research and Innovation Project under Grant XJ2020G060 and Grant XJ2019G039, and in part by the Xinjiang Uygur Autonomous Region Higher Education Innovation Project under Grant XJEDU2017T002.

ABSTRACT At present, occlusion and appearance similarity pose severe challenges to person re-identification tasks. Although many robust deep convolutional neural networks alleviate these problems, convolutional layers with limited receptive fields cannot model global semantic information well. In addition, in the person re-identification model, many metric losses ignore or destroy the intra-class structure of the sample, which makes the model difficult to be optimized. Therefore, we design a discriminative Re-identification model with global-local attention and adaptive weighted rank list loss (GLWR). Specifically, our global-local attention (GL-Attention) learns the semantic context in the channel and spatial dimensions. By learning the dependencies between features, GL-Attention integrates global semantic information into local features to extract discriminative features. Unlike rank list loss, our adaptive weighted rank list loss (WRLL) adaptively assigns weights according to the metric distance between the negative sample and the input image, which further improves the performance of the model. Experimental studies on three public datasets (Market-1501, DukeMTMC-ReID and CUHK03) indicate that the performance of our GLWR is significantly superior to many of the latest algorithms.

INDEX TERMS Person re-identification, deep learning, attention, loss function.

I. INTRODUCTION

The purpose of the person re-identification task (Re-id) is to retrieve a specific person through multiple non-overlapping cameras. The person may appear at different times or in different places. Re-id has broad application prospects, e.g., intelligent monitoring system and large-scale person tracking. Human identification and detection algorithms are usually used in the field of intelligent monitoring, but there are certain differences between them. The detection task [1]–[3] mainly focuses on the location and category of the target, while the human identification task mainly focuses on whether people with the same identity can be correctly identified in the gallery. In [4], [5], researchers use gait features and body movements to identify specific people. In order to enable the monitor to be used in the dark environment, the authors [46]–[48] use thermal imaging technology to identify human behavior. Compared with gait recognition and behavior recognition, the person re-identification only needs to extract the appearance features of the person, such as clothes, shoes, and bags. However, Re-id is a challenging task due to occlusion, similar appearance, size changes, etc. As shown in Fig. 1(a) and (b), the target is blocked by other people or objects while walking. In Fig. 1(c) and (d), some people are difficult to distinguish because of their similar appearance. From Fig. 1, we see that we see that BagTricks [14] cannot solve these problems well. To address these challenges effectively, we design a discriminative Re-id model.

In recent years, the Re-id algorithms [6], [7], [35] with deep learning have made tremendous progress in solving the aforementioned problems. However, the convolution kernel with a finite receptive field only extracts local features, which makes it difficult to learn global semantic information. To utilize rich global semantic information, Woo *et al.* [21] used a large convolution kernel to expand the receptive field. Lou *et al.* [43] found that the effective receptive field of the model is only a

The associate editor coordinating the review of this manuscript and approving it for publication was Gianluigi Ciocca.

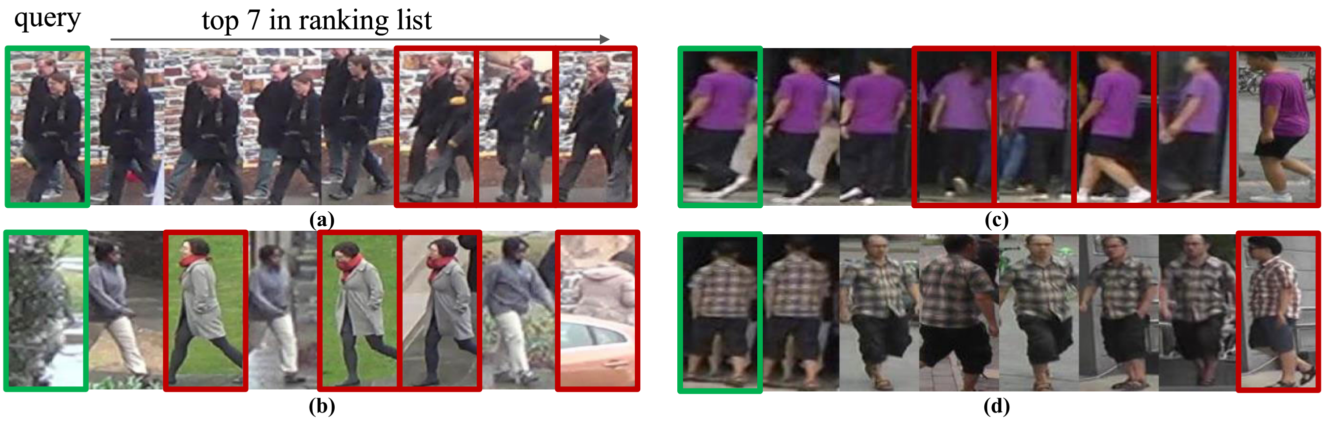


FIGURE 1. The ranking results of the BagTricks [9] on the DukeMTMC-ReID and Market-1501. The green boxes represent the query images, and the red boxes indicate the images with the wrong match.

small part of the theoretical receptive field. Therefore, these models cannot capture global context information effectively. In addition, the loss function is an essential part of supervised training, which is used to optimize model parameters. Generally, researchers [14]–[17] use the classification loss (ID loss) and metric losses to deal with Re-id tasks. ID loss aims at assigning people to their class correctly. The function of metric losses is to measure the similarity between people. However, Wang *et al.* [18] found that many metric losses ignore or destroy the intra-class structure to a certain extent, making it difficult for the model to extract features. In order to preserve the structure of the sample, Wang *et al.* [18] designed a rank list loss (RLL). When the measured distance between the input image and the hard negative sample is very close, the Re-id model mistakenly treats the negative sample as a positive sample. To solve this problem, RLL [18] manually adjusted the parameters to optimize the model, which makes it less flexible.

By analyzing the above problems, we design a discriminative Re-id model with global-local attention and adaptive weighted rank list loss (GLWR). In particular, GLWR is proposed on top of BagTricks [14]. Ours GL-Attention integrates global semantic information into local features to learn subtle distinguishing features. Specifically, the position attention module contains a global branch and a local branch. First, the global branch learns the dependence between arbitrary features through a self-attention mechanism that is not constrained by the receptive field. Second, the local branch learns the semantic relationship between local features by the convolutional block attention module (CBAM) [21]. Finally, the two branches are fused. The channel attention module also contains two branches, which are used to capture the correlation between channels and assign weights to each channel. From Fig. 2, we see that GL-Attention can focus on the discriminative human body regions. At the same time, we adopt an adaptive pooling layer [19] to capture the discriminant features of specific domains. Unlike RLL [18], we propose a WRLL, which is more flexible and adaptable. WRLL dynamically assigns weights according to the metric distance between the input image and the negative sample. Specifically, if the metric distance is too close, WRLL will assign a large weight to mine hard samples, which can prevent the Re-id model from treating negative samples as positive samples. The experimental studies illustrate that the mean average precision (mAP) of our GLWR is higher than BagTricks [14] by 22.3 % on the Cuhk-03 datasets.

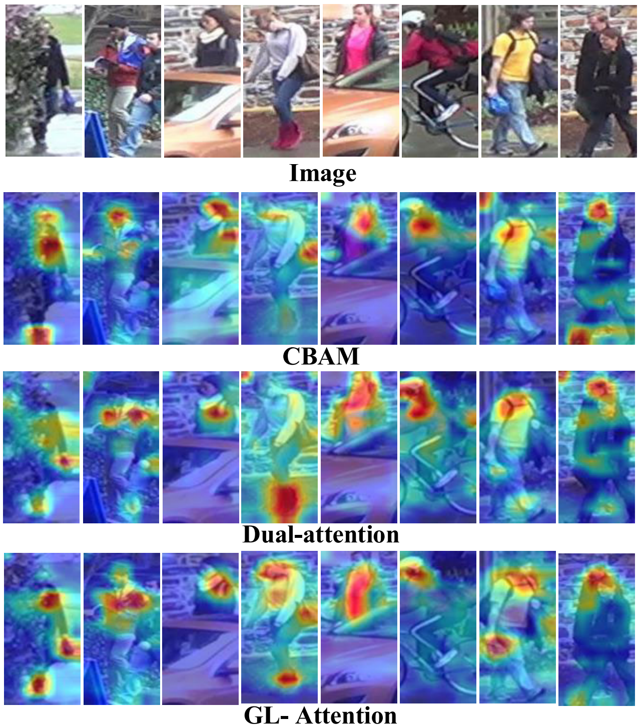


FIGURE 2. Heat map of Dual-attention [20], CBAM [21] and GL-Attention on DukeMTMC-ReID. CBAM makes the model extract features effectively. However, due to the limitation of the receptive field, it cannot capture some highly distinguishable features through global context information. For example, CBAM cannot pay attention to the bag held by the first and seventh people, but the GL-Attention can do it.

In summary, this study has the following contributions:

- We propose a global-local attention (GL–Attention), which integrates global semantic information into local features to extract discriminative features effectively.

- We propose an adaptive weighted rank list loss (WRLL) that, unlike Rank List Loss, is both adaptable and accurate.
- We adopt an adaptive pooling layer to capture the discriminant features of specific domains.
- Based on Bagtricks [14], we design a discriminative person re-identification model with global-local attention and adaptive weighted rank list loss (GLWR). We show experimentally that the proposed GLWR achieves the state - of - the - art results.

II. RELATED WORKS

The Re-id method based on deep learning includes two key parts: feature extractor and deep metric learning. In this section, we briefly introduce the work related to Re-id.

In order to train a powerful feature extractor to extract discriminant features, many classical methods [8]–[11], [29], [35] have been proposed. They trained discriminant models to extract features and used similarity measures to determine the similarity between images. Wang *et al.* [8] designed a multi-granularity network (MGN) that combines part multi-granularity information and global information to extract rich visual features. Zheng *et al.* [10] combined discriminant and generative learning to train powerful feature extractors. Although these models have achieved high performance, their structures are too complicated. To reduce the complexity of the model, most researchers [22]–[25], [29], [44], [33] skillfully use attention models to improve network performance. Tay *et al.* [22] presented an attribute attention network (AANet), which extracts the features of human key parts and integrates these features into the classification framework. Xia *et al.* [24] proposed a second-order-non-local attention (SONA), which uses second-order features to strengthen the connection between local features. Wang *et al.* [44] proposed a harmonious attention (HA), which can effectively learn soft pixel attention and hard region attention to extract features. Chen *et al.* [33] formulated a self-critical attention (SCAL), which can evaluate the quality of attention maps and provide strong supervisory signals to guide the model training. However, these attention models perform convolution operations under limited receptive fields, which cannot ensure that global context information is captured effectively. To solve this problem, the self-attention models [26]–[28] were proposed. Cao *et al.* [28] proposed a global context network (GcNet) to capture long-range dependencies, which is based on the non-local neural networks [27] (NL). Dual-Attention [20] learns global context information in spatial and channel dimensions, without being constrained by the receptive field. Although CBAM [21] cannot utilize global context information, it is a lightweight model and can focus on locally relevant features to extract salient features. In order to make the Re-id model integrate global semantic information into local features, we design a GL-Attention, which shows good performance.

In deep learning, the loss function is used as a supervisory signal to guide network training. Zheng *et al.* [30] designed an ID-discriminative embedding (IDE) that uses ID loss to optimize network parameters. They used the idea of classification to solve the Re-id challenge. If the model is trained only with ID loss, the performance of the model may be poor. The main reason is that the model needs an extra fully connected layer to predict the probability of people IDs during the training process. In the verification stage, the fully connected layer is removed and the features of the last pooling layer are used to measure the similarity between samples. To solve this problem, many researchers [12], [13] use multiple losses to optimize the network. The joint training strategy of triplet loss and ID loss is an effective method [31], [33], [34]. However, the triple loss is affected by the sample distribution, which leads to poor generalization ability of the network. The methods [35], [36] are proposed to improve the triplet loss, which has achieved good performance. However, in order to concentrate the positive samples within a certain range, they ignored the internal structure of the samples. Wang *et al.* [18] believed that the structural information of the sample is of great significance to the optimization of the model. Therefore, we proposed a WRLL based on RLL [18], which has higher flexibility and adaptability.

III. PROPOSED METHODS

In this section, we sequentially introduce the structure of the designed GLWR, the position attention model and channel attention model of GL-Attention, generalized-mean Pooling (GeM) [19], and adaptive weighted rank list loss (WRLL).

A. THE STRUCTURE OF GLWR

The proposed GLWR is a robust and simple Re-id model. In the Re-id task, the backbone network of many advanced models is ResNet-50 [37], which effectively solves the gradient explosion problem and improves the performance of the model by using shortcut connections. For the fairness of experimental comparison, we also use ResNet-50 as the backbone network of GLWR. Fig. 3 shows the structure of the GLWR, which has four important components: ResNet-50, GL-attention, GeM and WRLL. In Fig. 3, we add the designed GL-Attention after the first and third residual blocks, which can effectively extract different levels of discriminative features. In order to obtain a robust Re-id model, we adopt WRLL and ID loss to optimize the model parameters. Specifically, we add the WRLL after the GeM layer and place the ID loss behind the fully connected layer.

Table 1 describes the detailed information of the layer parameters in the proposed GLWR. We preprocess the original image and set the image size to 3 × 256 × 128, which is used as the input of GLWR. In Table 1, the size, stride, padding of Conv2d-1 are set as (7 × 7), (2 × 2) and (3 × 3), respectively. The size, stride, padding of Max pool are set as (3 × 3), (2 × 2) and (1 × 1), respectively. In ResNet-50 [37], He *et al.* have a detailed introduction to the structure of the Residual block-x, where x = {1, 2, 3, 4}. In order to describe the structure flow of GLWR concisely, we no longer introduce the structure of the Residual block-x.

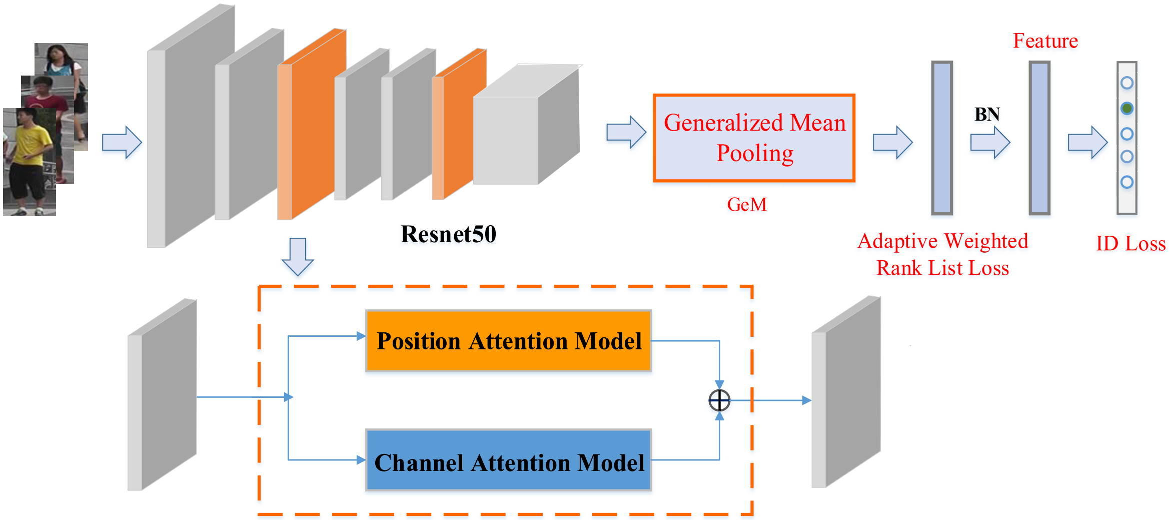


FIGURE 3. The Structure of GLWR.

TABLE 1. Description of the structure of the proposed GLWR.

Layer number	Layer name	Input size	Number of parameters	Output size	Layer connection (connected to)
0	Input layer	3×256×128		3×256×128	n/a
1	Conv2d-1	3×256×128	9,408	64×128×64	Input layer
2	BN		128		Conv2d_1
3	ReLu		0		BN
4	Max pool	64×128×64	0	64×64×32	ReLu
5	Residual block-1	64×64×32	232,256	256×64×32	Max pool
6	Position attention	256×64×32	132,454	256×64×32	Residual block_1
7	Channel attention	256×64×32	67,686	256×64×32	Residual block_1
8	add	256×64×32	0	256×64×32	Position attention & Channel attention
9	Residual block-2	256×64×32	12,19,584	512×32×16	add
10	Residual block-3	512×32×16	7,360,768	1024×16×8	Residual block_2
11	Position attention	1024×16×8	2,102,374	1024×16×8	Residual block_3
12	Channel attention	1024×16×8	1,056,870	1024×16×8	Residual block_3
13	add	1024×16×8	0	1024×16×8	Position attention & Channel attention
14	residual block-4	2048×16×8	14,964,736	2048×16×8	add
15	GeM	2048×16×8	0	2048	residual block_4
16	BN	2048×1×1	4,096	2048	GeM
17	Linear	2048	3,074,048	1051	BN
Total training parameters: 27,136,728					

B. Position Attention Model of GL-Attention

As shown in Fig. 4, we design a position attention model, which contains a global branch and a local branch. The global branch learns the spatial dependence between arbitrary features to capture contextual relationships. The local branch extracts local features, which makes the Re-id model focus on key information. We introduce the working principle of position attention through the following two steps.

In the first step, given the input feature $\boldsymbol{A} \in R^{C\times H\times W}$, the feature map $\boldsymbol{B} \in R^{\boldsymbol{C}\times \boldsymbol{H}\times \boldsymbol{W}}$ and the feature map $\boldsymbol{C} \in R^{\boldsymbol{C}\times \boldsymbol{H}\times \boldsymbol{W}}$ are generated by the convolution layer, respectively. Then, the size of $\boldsymbol{B}$ and $\boldsymbol{C}$ is adjusted to $R^{C\times N}$, where $H \times W = N$. Next, the matrix multiplication is carried out for the transformation of feature $\boldsymbol{C}$ and feature $\boldsymbol{B}$. Finally, we obtain the spatial attention map $\boldsymbol{S} \in R^{N\times N}$ by a softmax layer:

$$S_{ji} = \frac{\exp(B_i \otimes C_j)}{\sum_{i=1}^{N} \exp(B_i \otimes C_j)} \tag{1}$$

where S_{ji} is the effect of the i^{th} feature on the j^{th} feature. B_i and C_j are arbitrary in the feature map. The more similar they are, the more relevant they are.

In the second step, we feed the feature $\boldsymbol{A}$ into CBAM and the convolution layer respectively to obtain the feature $\boldsymbol{E} \in R^{C\times H\times W}$ and the feature $\boldsymbol{D} \in R^{C\times H\times W}$. Then, the matrix multiplication is carried out for the spatial attention $\boldsymbol{S}$ and the

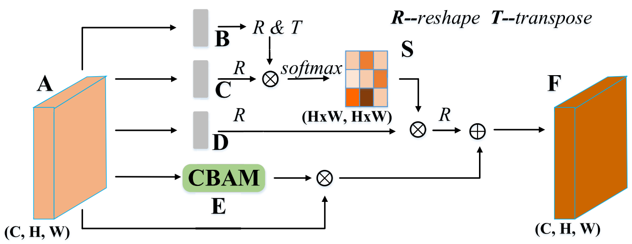


FIGURE 4. Position attention model.

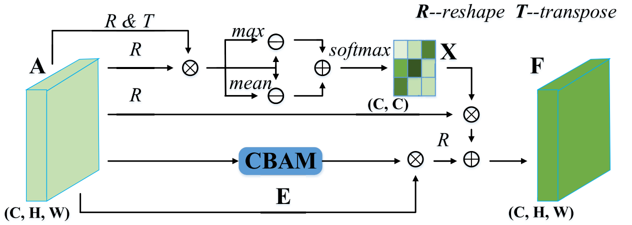


FIGURE 5. Channel attention mode.

feature $\boldsymbol{D}$. Next, the size of calculation results is adjusted to $R^{C\times H\times W}$, and the calculation results are multiplied by the parameter α. Finally, we perform element-wise summation on it and feature $\boldsymbol{E}$ to obtain the feature $\boldsymbol{F} \in R^{C\times H\times W}$:

$$F_j = \alpha \sum_{i=1}^{N} \left(S_{ji} \otimes D_i\right) + E_j \tag{2}$$

where α is an adaptive parameter [38]. Eq. 2 shows that the weighted sum of all position features and E_j is the j^{th} feature of $\boldsymbol{F}$. E_j is the weighted sum of local features, which can learn local features. Therefore, the position attention model not only uses the global context to capture the dependencies between global features, but also uses local information to capture the dependencies between neighborhood features.

C. CHANNEL ATTENTION MODEL OF GL-ATTENTION

As shown in Fig. 5, in order to capture the dependencies between channels, we design a channel attention model. We introduce the working principle of channel attention through the following two steps.

In the first step, given the input feature $\boldsymbol{A} \in R^{\boldsymbol{C}\times\boldsymbol{H}\times\boldsymbol{W}}$, and reshape $\boldsymbol{A}$ to $R^{C\times N}$. Then, the matrix multiplication is carried out for the transformation of feature $\boldsymbol{A}$ and feature $\boldsymbol{A}$. Next, the average and maximum value of each column of the matrix multiplication result are calculated to obtain two new feature maps $\{\boldsymbol{G}, \boldsymbol{H}\} \in R^{C\times C}$. Finally, we get the channel attention map $\boldsymbol{X} \in R^{C\times C}$ by a softmax layer:

$$G_{ij} = \max(A_i \otimes A_j) \tag{3}$$

$$H_{ij} = \text{mean}\left(A_i \otimes A_j\right) \tag{4}$$

$$X_{ji} = \frac{\exp(G_{ij} + H_{ij} - 2 \times (A_i \otimes A_j))}{\sum_{i=1}^{C} \exp(\{G_{ij} + H_{ij} - 2 \times (A_i \otimes A_j))} \tag{5}$$

where X_{ji} is the effect of i^{th} channel on j^{th} channel. We use the max function and the mean function to calculate the maximum value and average value of each column of $A_i \otimes A_j$, and the matric size of the result is adjusted to $R^{C\times C}$. It is worth noting that Dual-attention [20] is used for scene segmentation tasks, and its purpose is to assign edge information of different objects to its category accurately. However, in the Re-id task, we need to deal with occlusion and appearance similarity. Through experimental analysis, we find that if we only use the maximum function, some fine-grained features will be lost, so we add the average function.

In the second step, we feed feature $\boldsymbol{A}$ into CBAM to obtain a new feature map $\boldsymbol{E} \in R^{C\times H\times W}$.Then, $\boldsymbol{A}$ is adjusted to $R^{C\times N}$ and the matrix multiplication is carried out for the channel attention $\boldsymbol{X}$ and $\boldsymbol{A}$. Next, the calculation results are multiplied by the parameter β. Finally, we perform an element-wise summation operation on it and feature $\boldsymbol{E}$ to obtain the feature $\boldsymbol{F} \in R^{C\times H\times W}$:

$$F_j = \beta \sum_{i=1}^{N} \left(X_{ji} \otimes A_i\right) + E_j \tag{6}$$

with the training of the network, the parameter β can be updated gradually. In Eq. 6, the weighted sum of all channel features and E_j is the feature of each channel, so that the semantic relationship between the feature maps can be captured effectively.

D. GENERALIZED-MEAN (GEM) POOLING

In the Re-id task, we need to deal with many complex situations, such as occlusion, similar appearance, etc. The average pooling layer and the maximum pooling layer can retain the main features and eliminate redundant information, but they cannot catch the discriminative features in specific domains well. Therefore, we adopt an adaptive pooling layer, named generalized mean pooling (GeM) [19]. The formula is as follows:

$$\mathrm{f} = [f_1 \ldots f_k \ldots f_K]^T, f_k = \left(\frac{1}{|X_k|} \sum\nolimits_{x_i \in X_k} x_i^{p^k}\right)^{\frac{1}{p^k}} \tag{7}$$

where p^k is a hyper-parameter, which is learned during model training. When $p^k \rightarrow \infty$, the above formula approximates the maximum pooling. When $p^k \rightarrow 1$, the formula approximates the average pooling.

E. ADAPTIVE WEIGHTED RANK LIST LOSS

In this section, we focus on explaining the adaptive weighted rank list loss (WRLL), which preserves the structural information of the samples and learns a hypersphere for each class. WRLL can adaptively allocate parameters based on the hard samples of the dataset.

Let us first explain the symbols in the following formula. $X = \{(x_i, y_i)\}_{i=1}^{N}$ is the training set, where (x_i, y_i) is the i^{th} sample and the corresponding label. There are class C samples in the training set, namely $y_i \in [1, 2, \ldots, C]$. $\{(x_i^c)\}_{i=1}^{N_C}$ are all samples in the class C. $P_{c,i}^*$ is positive samples, $N_{c,i}^*$ is negative samples.

Given an image X_i, we aim to make it closer to its positive point and put all the positive points together to learn a hypersphere with radius $\alpha - m$. The formula is as follows:

$$L_m\left(x_i, x_j; f\right) = \left(1 - y_{ij}\right)|\alpha - d_{ij}| + y_{ij}|d_{ij} - (\alpha - m)| \quad (8)$$

$$L_p(x_i^c; f) = \frac{1}{\left|P^*_{c,i}\right|}\sum_{x_j^c \in P^*_{c,i}} L_m\left(x_i^c; x_j^c; f\right) \quad (9)$$

where $d_{ij} = \left\|f\left(x_i\right) - f(x_j)\right\|_2$ is the Euclidean distance between two samples. In order to make keep the input away from its negative point, its distance from its negative sample is at least α. The minimum distance between the positive and negative samples is m.

There are fewer positive samples than negative samples. To avoid the poor generalization ability of the network, we assign adaptive weights by using softmin weight distribution. The formula is as follows:

$$w_{ij} = \frac{\exp(-d_{ij}^n)}{\sum_{x_j^c \in N^*_{c,i}} \exp(-d_{ij}^n)} \quad (10)$$

where d_{ij}^n is the Euclidean distance between the negative sample and the input image. Eq. 10 shows that the adaptive weight mine the hard samples. When a few difficult samples appear in a batch, the adaptive weight dynamically assigns weights to the loss according to the distance between the two samples. In Eq. 11, the parameter T needs to be adjusted to make the model perform best on a dataset. However, when using other datasets, the parameter T needs to be readjusted. Therefore, our weighted strategy is more flexible and effective than RLL [18].

$$RLL_{w_{ij}} = \exp\left(T \cdot \left(\alpha - d_{ij}\right)\right), x_j^k \epsilon N^*_{c,i} \quad (11)$$

where T is the slope, which is used to adjust the amplitude of weight change. To keep the input away from its negative point, its distance from the negative sample is at least α.

Similarly, the negative sample loss function is as follows:

$$L_N(x_i^c; f) = \sum_{x_j^k \epsilon |N^*_{c,i|}} \frac{w_{ij}}{\sum_{x_j^k \epsilon |N^*_{c,i}|} w_{ij}} L_m(x_i^c; x_j^c; f) \quad (12)$$

In WRLL, we optimize the loss function of positive and negative samples jointly:

$$L_{WRLL}\left(x_i^c; f\right) = L_p\left(x_i^c; f\right) + L_N\left(x_i^c; f\right) \quad (13)$$

Finally, we adopt the multiple losses joint learning strategy as follows:

$$L_{Total} = L_{ID} + w * L_{WRLL} \quad (14)$$

where L_{ID} is the cross-entropy loss function. We need to fine-tune the weight w.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The proposed GLWR is implemented in Pytorch 0.4.1. We do a large number of experiments on the tesla v100 GPU with 16GB RAM.

TABLE 2. Detailed introduction to the Market-1501.

Subset	# Identities	# Images	# Cameras
Train	751	12936	6
Query	750	3368	6
Gallery	751	15912	6

TABLE 3. Detailed introduction to the DukeMTMC-ReID.

Subset	# Identities	# Images	# Cameras
Train	702	16522	8
Query	702	2228	8
Gallery	1110	17661	8

TABLE 4. Detailed introduction to the Cuhk-03 (Detected).

Subset	# Identities	# Images	# Cameras
Train	767	7356	2
Query	700	1400	2
Gallery	700	5332	2

A. DATASETS AND IMPLEMENTATION DETAILS

1) DATASETS

According to the standard practices, we conduct experimental studies on three authoritative datasets, namely Market-1501 [40], DukeMTMC-ReID [41] and the small–scale Cuhk-03(Detected) [6] datasets. Tables 2, 3, and 4 describe these three datasets in detail. The training set is used to train all models. The query set and gallery set are used to test all models. It is worth noting that the identities of the training images are different from those of the test images.

2) EVALUATION METRICS

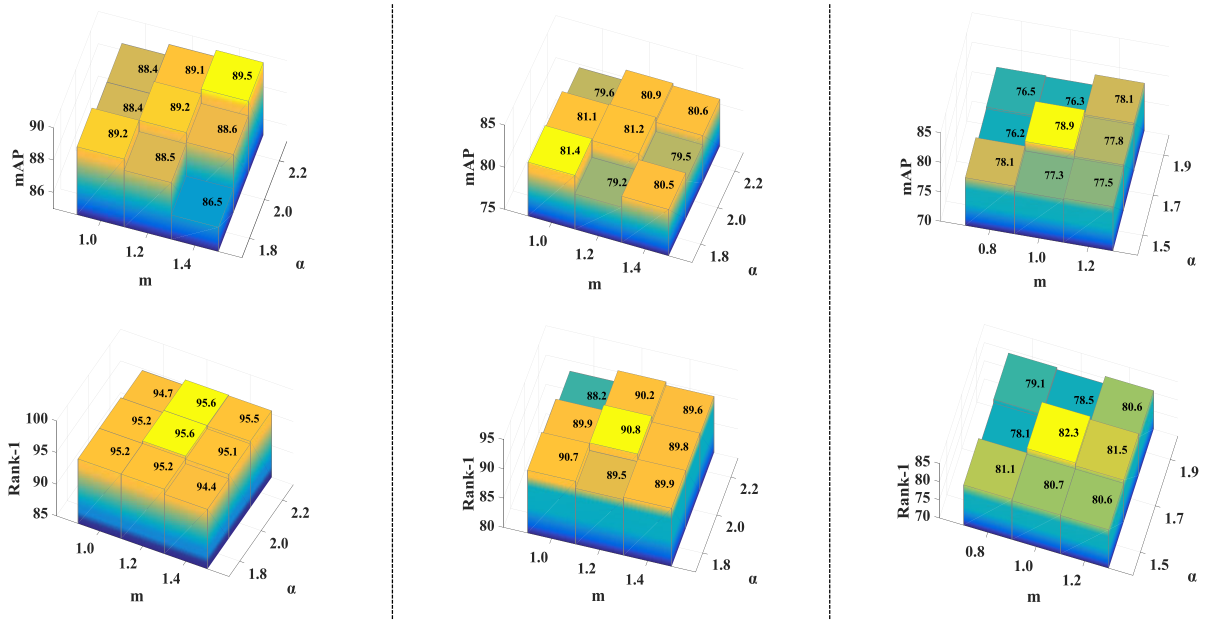
We adopt the cumulative match characteristic (CMC) [45] and the mean average precision (mAP) [40] to evaluate the performance of GLWR. CMC is a classic evaluation indicator in the Re-id tasks. The abscissa of the CMC curve is Rank-n, where n = 1, 3, 5, etc. The ordinate is the recognition accuracy. The CMC curve shows the recognition accuracy of Rank-n, which can effectively evaluate the performance of the model. The recognition accuracy of Rank-n represents the probability of finding the correct identity in the first n recognition results. Many researchers usually use Rank-1 to evaluate the performance of the model. The mAP represents the average accuracy of retrieving the specified identity correctly in the database, which can comprehensively measure the performance of the model.

3) DATA AUGMENTATION

We use random erasing [49], horizontal flipping and random cropping [50] to preprocess the images. The input images are resized to 256 × 128.

4) TRAINING AND SETTINGS

The backbone network is pre-trained on ImageNet [39]. We take Bagtricks [14], which only uses the classification loss, as our baseline network. Adam optimizer is used to train all models. Both the weight decay and the initial learning



(a) The Market-1501 dataset.

(b) The DukeMTMC-ReID dataset

(c) The Cuhk-03 dataset.

FIGURE 6. **Parameter optimization for GLWR in three dataset.**

rate are set to 0.0005. The parameter $w = 0.5$ in Eq. 13. The hyper-parameters α and m in the WRLL are analyzed in Section IV.B. The decay steps of the learning rate $lr(t)$ are as follows:

$$lr(t) = \begin{cases} 5.0 \times 10^{-4} \times \dfrac{t}{10}, & t \leq 10 \\ 1.0 \times 10^{-4}, & 10 < t \leq 50 \\ 2.0 \times 10^{-5}, & 50 < t \leq 90 \\ 4.0 \times 10^{-6}, & 90 < t \end{cases} \tag{15}$$

B. HYPER-PARAMETERS ANALYSIS

In Eq. 8, there are two hyper-parameters α and m, which have a great effect on the performance of the GLWR. The $(\alpha - m)$ denotes the radius of the hypersphere. The sample distribution of the dataset determines the size of the radius.

Improper adjustment of radius size can lead to poor performance. To obtain better performance, we conduct ablation studies on these two hyper-parameters. Inspired by the parameter adjustment of the RLL [18], we adopt the control variable method to fine-tune the parameters. For example, if we adjust the parameter α to the optimal value, we fix the parameter m, and so on. As shown in Fig. 6, we optimize the parameters α and m on the three datasets, respectively.

From Fig. 6, we see that parameters α and m are sensitive to the performance of the model. From the experimental results, we draw several conclusions. On the Market-1501 dataset, the optimal parameters are set as follows: $\alpha = 2.2$, $m = 1.4$. On the DukeMTMC-ReID dataset, the optimal parameters are set as follows: $\alpha = 1.8$, $m = 1.0$. On the Cuhk-03 dataset, the optimal parameters are set as follows: $\alpha = 1.7$, $m = 1.0$.

C. COMPARISON WITH THE STATE-OF-THE-ART

The methods of MGN [8], CAM [9], GD-Net [10], MSBA [12], BagTricks [14], AGW [15], ABD-Net [16], RAG-SC [17], AANet [22], SONA [24], RRGCCAN [26], IANet [29], EMM [32], MPM-LTL [35], HA-CNN [44] and SCAL [33] are selected for comparison. It is worth noting that we do not use the re-ranking strategy. The backbone network of these comparison methods is ResNet-50 [37]. Table 5, Table 6 and Table 7 indicate the performance comparisons between our GLWR and the state-of-the-art methods.

Table 5 shows that our GLWR achieves 89.5% mAP and 95.5% Rank-1 on the Market-1501. Compared with BagTricks [14], our GLWR increases mAP by 3.6% and Rank-1 by 1.0%, respectively. Compared with AGW [15] and MSBA [12], which are also designed based on BagTricks [14], our GLWR performs better than them. Table 6 shows that GLWR achieves 81.4% mAP and 90.7% Rank-1 on the DukeMTMC-ReID. Compared with BagTricks [14], our GLWR increases mAP by 5.0% and Rank-1 by 4.3%, respectively. Compared with AGW [15], our GLWR increases mAP by 1.8% and Rank-1 by 1.7%, respectively. Table 7 shows that our GLWR achieves 78.9% mAP and 82.3% Rank-1 on the Cuhk-03(Detected). Compared with BagTricks [14], our GLWR increases mAP by 22.3% and Rank-1 by 23.5%, respectively. Compared with AGW [15], our GLWR increases mAP by 16.9% and Rank-1 by 18.7%. The mAP of GLWR is 4.4% higher than the advanced method currently, RAG-SC [17].

On the Market-1501, DukeMTMC-ReID and CUHK-03, by comparing with other algorithms, GLWR achieves the best performance, which is 0.7%, 1.3%, and 1.6% higher than the second best method on mAP, respectively. Our GLWR has

TABLE 5. Comparison with the state-of-the-art methods on the Market-1501.

Method	mAP	Rank-1
AANet [22] (CVPR2019)	83.4	93.9
MSBA [12] (IEEE Access 2020)	88.4	95.5
IANet [29] (CVPR2019)	83.1	94.4
MPM-LTL [35] (IEEE Access 2020)	80.5	93.9
SONA [24] (CVPR2019)	88.8	95.6
AGW [15] (arXiv2020)	87.8	95.1
EMM [32] (IEEE Access 2020)	87.1	95.4
ABD-Net [16] (CVPR2019)	88.3	95.6
MGN [8] (ACM2018)	86.9	95.7
HA-CNN[44](CVPR2018)	75.7	91.2
RRGCCAN [26] (IEEE Access 2020)	84.8	91.2
CAM [9] (CVPR2019)	84.5	94.7
GD-Net [10] (CVPR2019)	86.0	94.8
BagTricks [14] (CVPR2019)	85.9	94.5
GLWR(Ours)	**89.5**	**95.5**

higher performance than the original BagTricks [14] under various indicators. The above experiments show that GLWR is a powerful Re-id model

D. ABLATION EXPERIMENTS

In this section, we do a lot of ablation experiments to analyze the effectiveness of GL-Attention, WRLL and GeM.

1) THE EFFECT OF GL-ATTENTION

To study the effectiveness of the GL-Attention, we add GL-Attention to the baseline network (B). We choose some advanced attention models for comparison, including CBAM [21], Dual-attention [20], NL [27] and GcNet [28]. For the fairness of comparison, we re-implement these attention models on top of the baseline. Fig. 7 and 8 indicate the experimental results of these attention models.

In Fig. 7 and 8, we observe that: 1) on the Market-1501, the performance of GL-Attention is higher than that of the baseline, 2.8% and 1.0% higher on mAP and Rank-1, respectively; 2) on the DukeMTMC-ReID, the performance of GL-Attention is significantly higher than that of the baseline, 3.0% and 3.0% higher on mAP and Rank-1, respectively; 3) compared to other attention models, the GL-Attention is significantly superior to them. Experiments indicate that our GL-Attention has a significant impact on the baseline performance.

2) THE EFFECT OF ADAPTIVE WEIGHTED RANK LIST LOSS

To study the effect of WRLL, we do some comparative experiments on three datesets. RLL [18] pointed out that when the parameter T in Eq. 11 is large, the performance of the network will decline. According to the previous work, we set the parameter T to 3. Table 8 and 9 indicate the comparison results of the RLL and the WRLL.

TABLE 6. Comparison with the state-of-the-art methods on the DukeMTMC-ReID.

Method	mAP	Rank-1
AANet [22] (CVPR2019)	74.3	87.7
MSBA [12] (IEEE Access 2020)	80.1	90.1
IANet [29] (CVPR2019)	73.4	83.1
MPM-LTL [35] (IEEE Access 2020)	67.1	83.4
SONA [24] (CVPR2019)	78.3	89.5
AGW [15] (arXiv2020)	79.6	89.0
EMM [32] (IEEE Access 2020)	78.8	89.9
ABD-Net [16] (CVPR2019)	78.6	89.0
MGN [8] (ACM2018)	78.4	88.7
HA-CNN[44](CVPR2018)	63.8	80.5
RRGCCAN [26] (IEEE Access 2020)	77.7	86.0
CAM [9] (CVPR2019)	72.9	85.8
GD-Net [10] (CVPR2019)	74.8	86.6
BagTricks [14] (CVPR2019)	76.4	86.4
SCAL(spatial)[33](ICCV2019)	79.6	89.0
SCAL(channel)[33](ICCV2019)	79.1	88.9
GLWR(Ours)	**81.4**	**90.7**

TABLE 7. Comparison with the state-of-the-art methods on the CUHK-03.

Method	mAP	Rank-1
MSBA [12] (IEEE Access 2020)	72.9	76.2
MPM-LTL [35] (IEEE Access 2020)	62.7	56.6
SONA [24] (CVPR2019)	77.3	79.9
AGW [15] (arXiv2020)	62.0	63.6
EMM [32] (IEEE Access 2020)	67.3	69.9
RAG-SC [17] (CVPR2020)	74.5	79.6
MGN [8] (ACM2018)	66.0	68.0
HA-CNN[44](CVPR2018)	38.6	41.7
CAM [9] (CVPR2019)	64.2	66.6
BagTricks [14] (CVPR2019)	56.6	58.8
SCAL(channel)[33](ICCV2019)	68.6	71.1
GLWR(Ours)	**78.9**	**82.3**

In Table 4 and 5, we observe that: 1) our WRLL enhances the baseline performance significantly. In particular, on the Cuhk-03, WRLL improves the mAP of the baseline by 19.5%. This is because WRLL learns a hypersphere for each

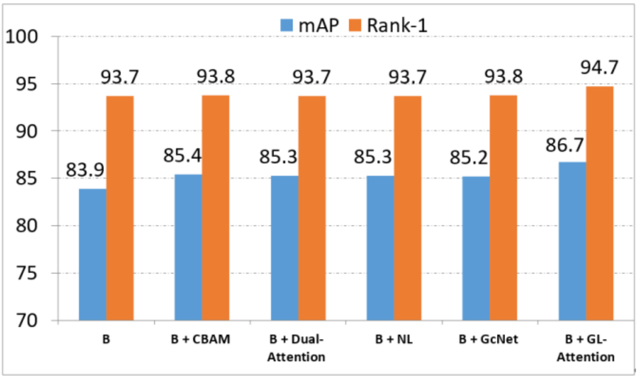


FIGURE 7. The mAP and the Rank-1 of CMC are used to evaluate the effectiveness of GL attention on market-1501.

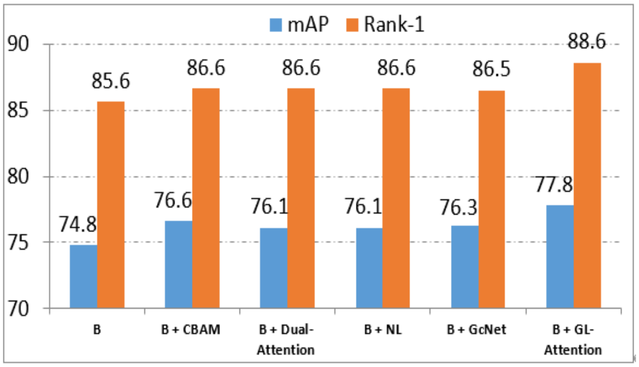


FIGURE 8. The mAP and the Rank-1 of CMC are used to evaluate the effectiveness of GL attention on the DukeMTMC-ReID.

TABLE 8. Evaluate the effectiveness of WRLL on the Market-1501 and DukeMTMC-ReID.

	Market-1501		DukeMTMC-ReID	
Method	mAP	Rank-1	mAP	Rank-1
B	83.9	93.7	74.8	85.6
B+RLL	86.3	94.4	76.8	87.8
B+WRLL	**87.7**	**94.9**	**78.0**	**89.0**

TABLE 9. Evaluate the effectiveness of WRLL on the CUHK-03.

	Cuhk-03	
Method	mAP	Rank-1
B	55.0	59.3
B+RLL	72.4	74.8
B+WRLL	**74.5**	**77.1**

class to protect the structure of the samples, which allows the baseline to capture more features in the dataset with fewer samples; 2) compared to RLL [18], our WRLL is accurate and flexible. On the cuhk-03, the performance of WRLL is better than that of RLL, 2.1% and 2.3% higher on mAP and Rank-1, respectively.

From Fig. 9, we observe that the loss curve decreases steadily, which means that our network is stable. Since the learning rate of the epoch 50 changes in Eq.15, the loss curve drops significantly at the epoch 50.

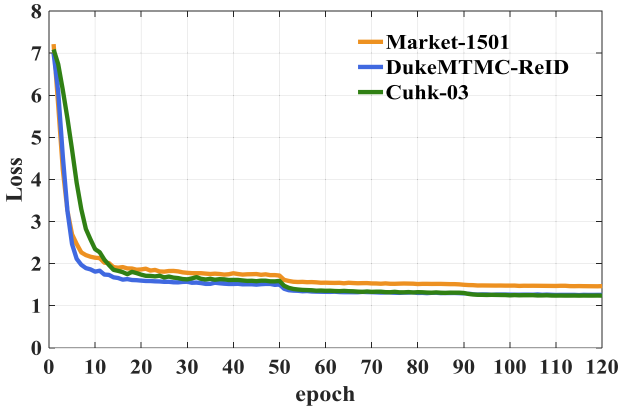


FIGURE 9. The loss curve of the B + WRLL on the three datasets.

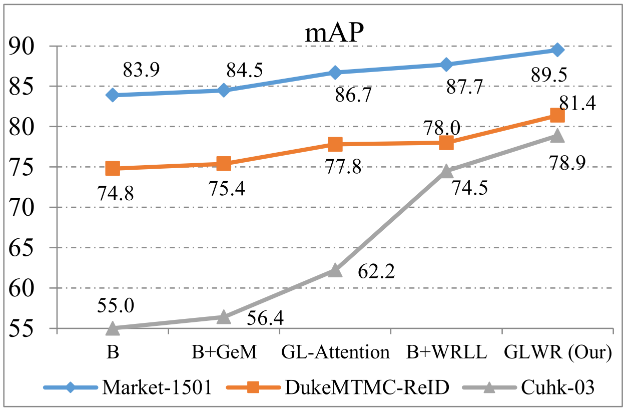


FIGURE 10. Evaluate the mAP of GLWR on the three datasets.

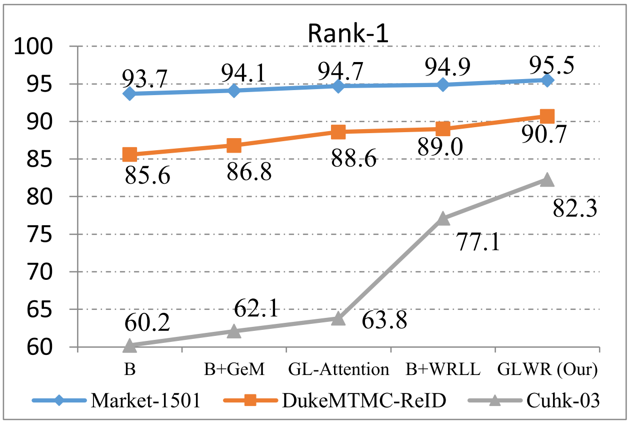


FIGURE 11. The performance of GLWR is evaluated by the rank-1 of CMC on the three datasets.

3) EVALUATE THE PERFORMANCE OF GLWR

We study the overall performance of GLWR. In Fig. 10 and 11, we evaluate the effects of the GeM, GL-Attention and WRLL on baseline performance, respectively.

In Fig. 10 and 11, we see that the introduction of each block significantly improves the performance of the baseline. From the above experimental results, we can draw the following conclusions: 1) our GLWR performs very well on the Cuhk-03 dataset with relatively small samples. On the Cuhk-03,

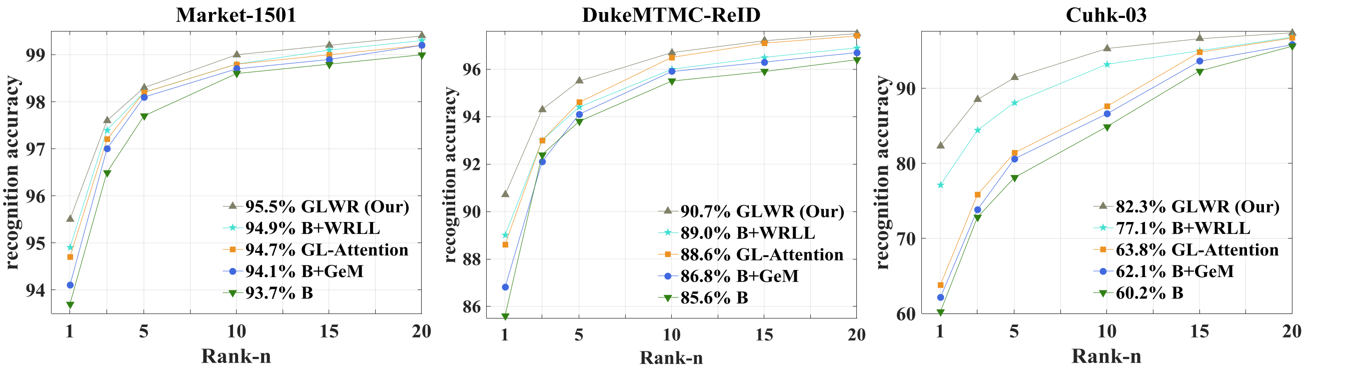


FIGURE 12. The CMC curve of the GLWR on the three datasets.

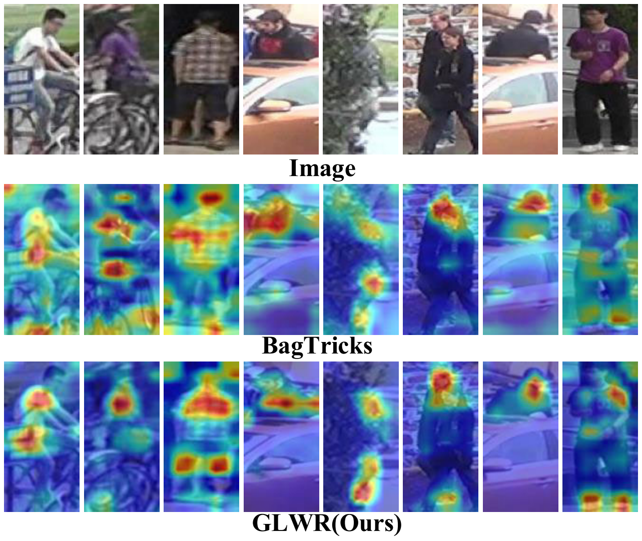


FIGURE 13. Heat map: BagTricks vs. GLWR(Ours).

the performance of our GLWR is significantly higher than that of the baseline, 23.9% and 22.1% higher on mAP and Rank-1, respectively; 2) the proposed GLWR is a powerful and simple Re-id model and every component of our GLWR is effective.

Fig. 12 shows the CMC curve of the GLWR on the three datasets. From Fig. 12, we see that the GeM, GL-Attention and WRLL improve the performance of the network respectively, and the fusion model of these three blocks can improve the network to the greatest extent.

E. VISUALIZATION OF RESULTS

We use the Grad-CAM [42] tool to analyze BagTricks [14] and GLWR. The Grad-CAM tool marks areas that the model considers important. The redder the marked area is, the more important it is. As shown in Fig. 13, we see that GLWR effectively captures the discriminant features of pedestrians and pays little attention to irrelevant information around pedestrians. From Fig. 2, comparing with other attention models, the GL-Attention clearly focuses on the highly discriminated information of pedestrians, which benefits from the model's fusion of global semantic information and local features.

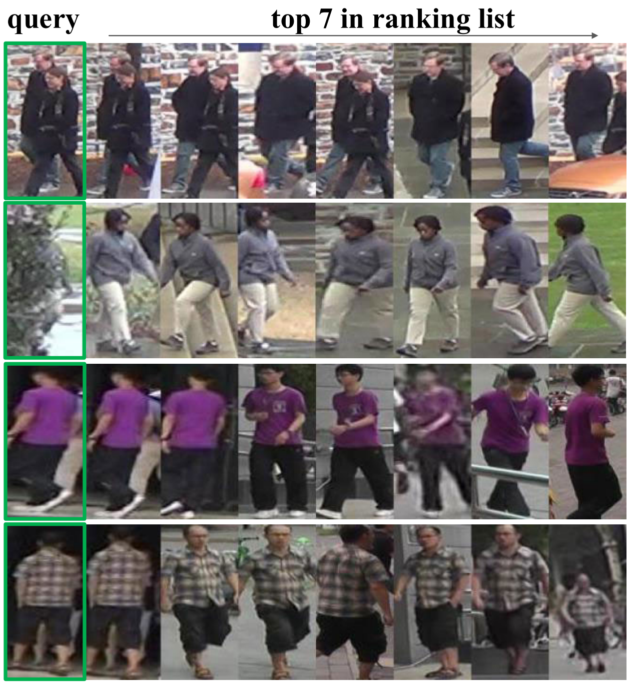


FIGURE 14. The ranking results of the proposed GLWR.

In Fig. 1, we see that there are many errors in the ranking results of BagTricks [14] due to the similar appearance and occlusion. We visualize the ranking results of GLWR. As shown in Fig. 14, we observe that compared to BagTricks [14], our GLWR has no matching errors, which indicates that our methods solves the problems of occlusion and similar appearance to a certain extent.

V. CONCLUSION

In this paper, we design a discriminative Re-id model with global-local attention and adaptive weighted rank list loss (GLWR) to solve occlusion and similar appearance problems in the Re-id task. Compared with other attention models, our GL-Attention contains global and local branches in the spatial dimension and channel dimension, respectively. It learns global context information and the dependency between neighborhood features, which can effectively extract discriminant features. It should be noted that the

GL-Attention can be embedded in any neural network, which can effectively enhance the ability of feature representation. At the same time, GLWR uses the adaptive pool layer to extract the discriminant features of specific domains. Our adaptive weighted rank list loss (WRLL) has more flexibility and adaptability than the Rank List Loss, and it can adaptively assign weights based on the measured distance between samples. Ablation studies on three authoritative datasets indicate that each component of GLWR is valid and the performance of GLWR is significantly higher than the existing methods. Especially, our GLWR is a simple and effective model that performs better on the dataset with relatively few samples.

In future research, we will focus on the generalization ability of the network, which can improve the performance of the model in cross-domain person re-identification. In addition, in order to make the model used in the dark environment, inspired by behavior recognition and gait recognition, we will try to use thermal imaging technology to assist in identifying people.

REFERENCES

[1] E. Jeon, J. Kim, H. Hong, G. Batchuluun, and K. Park, "Human detection based on the generation of a background image and fuzzy system by using a thermal camera," *Sensors*, vol. 16, no. 4, p. 453, Mar. 2016, doi: 10.3390/s16040453.

[2] N. Q. Truong, Y. W. Lee, M. Owais, D. T. Nguyen, G. Batchuluun, T. D. Pham, and K. R. Park, "SlimDeblurGAN-based motion deblurring and marker detection for autonomous drone landing," *Sensors*, vol. 20, no. 14, p. 3918, Jul. 2020, doi: 10.3390/s20143918.

[3] J. H. Kim, G. Batchuluun, and K. R. Park, "Pedestrian detection based on faster R-CNN in nighttime by fusing deep convolutional features of successive images," *Expert Syst. Appl.*, vol. 114, pp. 15–33, Dec. 2018, doi: 10.1016/j.eswa.2018.07.020.

[4] G. Batchuluun, H. S. Yoon, J. K. Kang, and K. R. Park, "Gait-based human identification by combining shallow convolutional neural network-stacked long short-term memory and deep convolutional neural network," *IEEE Access*, vol. 6, pp. 63164–63186, 2018, doi: 10.1109/ACCESS.2018.2876890.

[5] G. Batchuluun, R. A. Naqvi, W. Kim, and K. R. Park, "Body-movement-based human identification using convolutional neural network," *Expert Syst. Appl.*, vol. 101, pp. 56–77, Jul. 2018, doi: 10.1016/j.eswa.2018.02.016.

[6] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 152–159, doi: 10.1109/CVPR.2014.27.

[7] H. Zhang, T. Si, Z. Zhang, R. Zhang, H. Ma, and S. Liu, "Local heterogeneous features for person re-identification in harsh environments," *IEEE Access*, vol. 8, pp. 83685–83692, 2020, doi: 10.1109/ACCESS.2020.2991838.

[8] G. Wang, Y. Yuan, J. Li, S. Ge, and X. Zhou, "Receptive multi-granularity representation for person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 6096–6109, 2020, doi: 10.1109/TIP.2020.2986878.

[9] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang, "Towards rich feature discovery with class activation maps augmentation for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 1389–1398, doi: 10.1109/CVPR.2019.00148.

[10] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 2133–2142, doi: 10.1109/CVPR.2019.00224.

[11] Y. Lu, Y. Wu, B. Liu, T. Zhang, B. Li, Q. Chu, and N. Yu, "Cross-modality person re-identification with shared-specific feature transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 13376–13386, doi: 10.1109/CVPR42600.2020.01339.

[12] H. Tan, H. Xiao, X. Zhang, B. Dai, S. Lai, Y. Liu, and M. Zhang, "MSBA: Multiple scales, branches and attention network with bag of tricks for person re-identification," *IEEE Access*, vol. 8, pp. 63632–63642, 2020, doi: 10.1109/ACCESS.2020.2984915.

[13] C. Dai, J. Feng, and R. Zhou, "Learning domain-specific features from general features for person re-identification," *IEEE Access*, vol. 8, pp. 155389–155398, 2020, doi: 10.1109/ACCESS.2020.3018627.

[14] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2597–2609, Oct. 2020, doi: 10.1109/TMM.2019.2958756.

[15] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," 2020, *arXiv:2001.04193*. [Online]. Available: http://arxiv.org/abs/2001.04193

[16] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang, "ABD-net: Attentive but diverse person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 8350–8360, doi: 10.1109/ICCV.2019.00844.

[17] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 3183–3192, doi: 10.1109/CVPR42600.2020.00325.

[18] X. Wang, Y. Hua, E. Kodirov, G. Hu, R. Garnier, and N. M. Robertson, "Ranked list loss for deep metric learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 5202–5211, doi: 10.1109/CVPR.2019.00535.

[19] F. Radenovic, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1655–1668, Jul. 2019, doi: 10.1109/TPAMI.2018.2846566.

[20] J. Fu, J. Liu, J. Jiang, Y. Li, Y. Bao, and H. Lu, "Scene segmentation with dual relation-aware attention network," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Aug. 3, 2020, doi: 10.1109/TNNLS.2020.3006524.

[21] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.

[22] C.-P. Tay, S. Roy, and K.-H. Yap, "AANet: Attribute attention network for person re-identifications," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 7127–7136, doi: 10.1109/CVPR.2019.00730.

[23] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," in *Proc. ECCV*, Munich, Germany, Sep. 2018, pp. 8–14, doi: 10.1007/978-3-030-01225-0_23.

[24] B. Bryan, Y. Gong, Y. Zhang, and C. Poellabauer, "Second-order non-local attention networks for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 3759–3768, doi: 10.1109/ICCV.2019.00386.

[25] B. Chen, W. Deng, and J. Hu, "Mixed high-order attention network for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 371–381, doi: 10.1109/ICCV.2019.00046.

[26] X. Chen, L. Zheng, C. Zhao, Q. Wang, and M. Li, "RRGCCAN: Re-ranking via graph convolution channel attention network for person re-identification," *IEEE Access*, vol. 8, pp. 131352–131360, 2020, doi: 10.1109/ACCESS.2020.3009653.

[27] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7794–7803, doi: 10.1109/CVPR.2018.00813.

[28] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Seoul, South Korea, Oct. 2019, pp. 1971–1980, doi: 10.1109/ICCVW.2019.00246.

[29] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 9309–9318, doi: 10.1109/CVPR.2019.00954.

[30] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned CNN embedding for person reidentification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 1, pp. 1–20, Jan. 2018.

[31] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017, doi: 10.1109/TIP.2017.2700762.

[32] F. Zhou, W. Chen, and Y. Xiao, "Deep learning research with an expectation-maximization model for person re-identification," *IEEE Access*, vol. 8, pp. 157762–157772, 2020, doi: 10.1109/ACCESS.2020.3019100.

[33] G. Chen, C. Lin, L. Ren, J. Lu, and J. Zhou, "Self-critical attention learning for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 9636–9645, doi: 10.1109/ICCV.2019.00973.

[34] S. Zhou, F. Wang, Z. Huang, and J. Wang, "Discriminative feature learning with consistent attention regularization for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 8039–8048, doi: 10.1109/ICCV.2019.00813.

[35] F. Zheng, T. Cai, Y. Wang, C. Deng, Z. Chen, and H. Zhu, "A mask-pooling model with local-level triplet loss for person re-identification," *IEEE Access*, vol. 8, pp. 138191–138202, 2020, doi: 10.1109/ACCESS.2020.3011961.

[36] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 1179–1188, doi: 10.1109/CVPR.2018.00129.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[38] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2019, pp. 7354–7363.

[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[40] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1116–1124, doi: 10.1109/ICCV.2015.133.

[41] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. ECCV*, Amsterdam, The Netherlands, 2016, pp. 17–35, doi: 10.1007/978-3-319-48881-3_2.

[42] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[43] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 8, Dec. 2016, pp. 4898-4906.

[44] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 2285–2294, doi: 10.1109/CVPR.2018.00243.

[45] R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior, "The relation between the ROC curve and the CMC," in *Proc. 4th IEEE Workshop Autom. Identificat. Adv. Technol. (AutoID)*, Buffalo, NY, USA, Oct. 2005, pp. 15–20, doi: 10.1109/AUTOID.2005.48.

[46] G. Batchuluun, Y. Kim, J. Kim, H. Hong, and K. Park, "Robust behavior recognition in intelligent surveillance environments," *Sensors*, vol. 16, no. 7, p. 1010, Jun. 2016, doi: 10.3390/s16071010.

[47] G. Batchuluun, D. T. Nguyen, T. D. Pham, C. Park, and K. R. Park, "Action recognition from thermal videos," *IEEE Access*, vol. 7, pp. 103893–103917, 2019, doi: 10.1109/ACCESS.2019.2931804.

[48] G. Batchuluun, J. H. Kim, H. G. Hong, J. K. Kang, and K. R. Park, "Fuzzy system based human behavior recognition by combining behavior prediction and recognition," *Expert Syst. Appl.*, vol. 81, pp. 108–133, Sep. 2017, doi: 10.1016/j.eswa.2017.03.052.

[49] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI*, New York, NY, USA, 2020, pp. 13001–13008, doi: 10.1609/aaai.v34i07.7000.

[50] Y. Wang, L. Wang, Y. You, X. Zou, V. Chen, S. Li, G. Huang, B. Hariharan, and K. Q. Weinberger, "Resource aware person re-identification across multiple resolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 8042–8051, doi: 10.1109/CVPR.2018.00839.

YONGCHANG GONG received the bachelor's degree from Yili Normal University, China, in 2018. He is currently pursuing the master's degree with the College of Information Science and Engineering, Xinjiang University, China. His research areas are computer vision and image processing.

LIEJUN WANG received the Ph.D. degree from the School of Information and Communication Engineering, Xi'an Jiaotong University, in 2012. He is currently a Professor with the School of Information Science and Engineering, Xinjiang University. His research interests include wireless sensor networks, encryption algorithm, and image intelligent processing.

YONGMING LI received the master's degree from Xinjiang University, China, in 2008. He is currently an Associate Professor with Xinjiang University. His current research interests include computer vision, video image processing, wireless sensor networks, and natural language processing.

ANYU DU received the master's degree from Xinjiang University, China, in 2016. She has a deep interest in the wireless sensor networks, digital image processing, computer vision, and deep hashing algorithms.