A Supervisory Mask Attentional Network for Person Re-Identification in Uniform Dress Scenes

1st Bo Li
Glodon AI Research.Glodon
Glodon
Beijing, China
lib-y@glodon.com

2th Ling Bai
Glodon AI Research.Glodon
Glodon
Beijing, China
bail-c@glodon.com

3th Yang Wang
Glodon AI Research.Glodon
Glodon
Beijing, China
wangy-dd@glodon.com

4th Zhe Wu
Glodon AI Research.Glodon
Glodon
Beijing, China
wuz-e@glodon.com

5th Tong Lin*
The Key Laboratory of Machine Perception
(MOE), School of EECS, Peking University,
Beijing 100871; Peng Cheng Laboratory,
Shenzhen 518052, China
lintong@pku.edu.cn

***Abstract*—**Person re-identification (Re-ID) aims at retrieving a person's identification across multiple non-overlapping cameras. While recent Re-ID methods have achieved significant success on a number of benchmark datasets, most of them are still insufficient in the scenes with unified or highly similar dressing like construction sites, schools and factories. To address this problem, we propose a supervisory mask attention network (SMA-Net). Our approach combines two key components: (1) ROI mask mapping (RMM) is a supervisory branch to provide ROI mappings that divide a person region into several parts; (2) Partial mask attention (PMA) integrates channel and space attention mechanisms that focus on local features and different accessories in each ROI. Therefore, the network can pay more attention to the local features and different accessories herein. Compared with the state of the art methods, SMA-Net demonstrates excellent performance on our dataset of construction scenes, with improvement of 6.65% in mAP and 3.8% in Rank-1 accuracy.

***Keywords*—**Person Re-identification, Partial Mask Attention, ROI Mask Mapping

I. INTRODUCTION

Person re-identification (Re-ID) is the process of associating images or videos of the same person taken from different angles and cameras. The key is to find discriminant features that represent a person. Many of the recent methods use deep learning models to extract features and achieve good performance on benchmark datasets. However, scenarios of similar clothes are relatively common in practical applications, such as schools, construction sites and factories, where uniform dress needs to be followed. Although the local feature learning approach can focus on the semantic information of local components on benchmark datasets such as Market1501 [1], the performance deteriorates evidently on the site person dataset as shown in Fig. 1.

The difficulty is that most people on construction sites dress very similar, with only subtle differences in sleeves, body contour and faces.The problem of Re-ID in similar dressing scenarios requires a more refined and specific model to provide strong identification ability for local features. To solve this problem, we need to answer: How to find out the similar parts of person, such as safety helmet, reflective clothing, mask, etc? How to reduce the attention to similar parts? How to improve the identification ability of different local special features? And how to control the amount of computation and complete the end-to-end training? In the literature, we find that there is no specific solution for uniform dress and accessories in the mainstream person re-identification methods.

Fig.1. Example images of several workers from our constructed dataset GC-2245.

To solve these problems, we propose an end-to-end supervisory mask attention network (SMA-Net), which focuses on more distinguishable feature regions. There are two parts in the network structure: (1) ROI mask mapping (RMM) is the supervised branch that provides a mapping of interesting regions. (2) Partial Mask Attention (PMA) integrates the mapping information in the supervised branch and performs masking operation, while adding channel and spatial attention mechanisms to the local feature extraction branch to enhance detailed feature attention. We adopt a multi-task joint learning network architecture that combines multi-granularity feature extraction network and local ROI feature supervisory branch. We employ object detection and semantic segmentation methods to find different masks of ROIs. We filter out the irrelevant features by reducing attention to similar dressing, so that the Re-ID model can learn more expressive distinguished features.

Due to the lack of a suitable datasets in the related research field, we build a similarly-dressed person re-identification dataset named as GC-2245 to facilitate further research.

This work was supported by NSFC Tianyuan Fund for Mathematics (No. 12026606), and National Key R&D Program of China (No. 2018AAA0100300). Tong Lin* is the corresponding author.

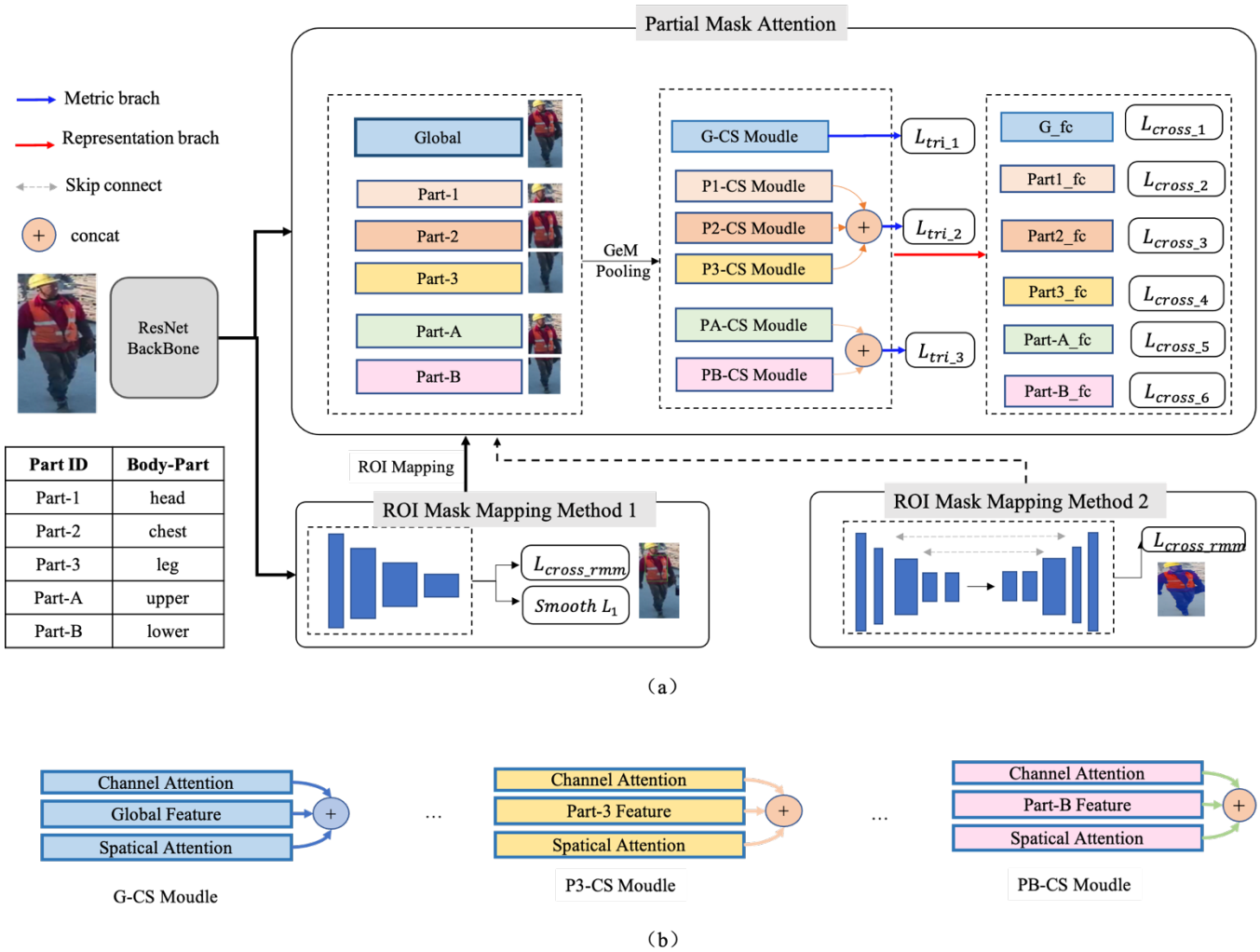


Fig.2. Workflow of SMA-Net. The overall architecture contains three parts. (a) PMA extracts multi-granularity features from person detection box and fuse feature internally. Features are divided into three levels: global features, head, chest and leg features and the upper and lower part of the body features. Combined with channel and spatial attention mechanisms, PMA module is to optimize learning loss and classification loss. RMM provides supervisory feature mapping for PMA. ROI Mask Mapping Method1 and ROI Mask Mapping Method2 demonstrate supervisory feature mapping methods based on target detection and image semantic segmentation. All above loss will be trained jointly. The internal structure of the attention fusion module in (a) is shown in (b).

The main contributions of this paper can be summarized as follows:

- An end-to-end feature extraction network SMA-Net is proposed in order to solve the similarly-dressed person re-identification problem, and achieve the state-of-the-art performance in this scenario.
- We propose ROI mask mapping (RMM), a module to extract specific fine-grained person Re-ID features. RMM module can be integrated into most Re-ID frameworks and trained effectively.
- We integrate object detection and semantic segmentation into the Re-ID task, which provides a new paradigm for multi-task learning.

II. Related Work

Person Re-ID technology can be divided into four categories.

(1) The first type is pose-guided method Spindle-Net [2]. It aligns the key points extracted from a pose estimator model to get ROI and then local features are extracted and multi-scale features are fused. But the pose estimator model may fail when objects are too small.

(2) The second type of Re-ID method is based on learning local features, such as Aligned-ReID [3], PCB [4], and MGN [5]. These models attempt to divide the imput image into multiple parts such that feature alignment and combination can be carried out. However, this line of methods can not diminish the importance of same dressing accessories.

(3) The third method is attention-based learning [6,7,8], such as ABD-Net [9] and ResNet-RGA [10], which enhances the discrimination power of foreground and background through channel attention, spatial attention, and global and local attention. However, the ability of local feature extraction often is insufficient.

(4) The fourth category is based on the generative adversarial networks [11-17], such as DG-Net [17]. The identity-related and identity-unrelated information can be decoupled by training encoders, such that identity-related features can be reserved in feature representations. Such methods are more suitable for domain adaptation where previous well-learned re-identification networks can be

transfered to a new task domain. It appears that GAN-based methods can not be applied to similar dressing problems in construction site scenes.

Our method mainly follows the second category. Compared with other methods, we propose RMM, a supervisory branch to reduce the attention on the same clothing, and PMA, a local feature attention module to enhance the ability to distinguish foreground and background.

III. Our Approach

The proposed SMA-Net structure is shown in as Fig. 2. The main idea is to reduce the attention contribution to the same dressed parts of person. SMA-Net is composed of two parts. Our method uses ResNet50 (with 50 layers of residual blocks) as backbone and shares weights among two branches.

A. Roi Mask Mapping (RMM)

RMM module is composed of object detection or semantic segmentation algorithms, aiming to provide area weights to the PMA module. Thus the extracted features focus on regions with large dressing differences and are more discriminant.

Specifically, there are two methods. The first method regards the task as object detection which has a similar structure to SSD [20]. Three detection branches are adopted to effectively decrease the calculation cost. The second method regards the task as image semantic segmentation which has a similar structure to U-Net [21]. We make some improvements to U-Net by adopting the ResNet50 convolution neural network for subsampling. This modification has following benefits:

- A deeper network improve segmentation accuracy;
- More shortcut connections are adds to the network, so that it can be better combined with the background semantic information for multi-scale segmentation;
- The ResNet structure is easy to be trained, which effectively prevents model degradation, gradient vanishing, and loss non-convergence;

B. Partial Mask Attention (PMA)

The PMA module adopts a multi-granularity feature extraction structure similar to MGN [5]. It provides a feature mapping of the region of interest in the local features through the object location or pixel-level classification information provided by RMM. At the same time, channel attention and spatial attention are integrated for the feature combination to enhance the expressive ability of local features. Different mask rate and mask mechanism will also affect the final performance of the algorithm, which will be described as follows.

The PMA consists of three network branches as shown in Fig. 2. The first branch is mainly responsible for extracting global features. Firstly, the dimensions of the features are reduced by global maximum pooling. At the same time, the global feature layer, channel attention global feature layer and spatial attention global feature layer are concatenated together for metric learning calculation. The classification is directly carried out by combining features with global connection layer.

The second branch extracts the features of the head, chest and legs of person. Unlike the global feature processing. In addition to the attention concatenating of each part, all the components need to be concatenated again for the metric learning.

The third branch is used to extract features from the upper and lower halves of people in the same way as the second branch. Three branches of feature extraction are computed as Equation (1) :

$$A_j = \sum_{i=1}^{i=j} S_i * g(x_i) \oplus g(x_i) \oplus C_i * g(x_i), j = 1,2,3 . \quad (1)$$

Where A_j represents the j-th sub-branch fusion feature. $\sum$ represents the fusion of features extracted from body parts in each sub-branch. S_i represents spatial attention mask matrix. C_i represents channel attention mask matrix, x_i and $g(x_i)$ is generic representation of feature extraction function. $*$ represents the matrix multiplication, and $\oplus$ represents the matrix concatenation. Particularly, A_1 represents the feature output results of the combination of global branches, A_2 represents the feature output results of the combination of upper and lower halves, and A_3 represents the feature output results of the combination of head, chest, leg.

C. Loss Functions

Three kinds of loss functions are included in the SMA-Net. The triplet loss [22] is used in the PMA's three feature extraction branches to measure loss :

$$L_{tri} = max\,(d(a,p) - d(a,n) + margin, 0). \quad (2)$$

Here the triplet loss first builds a ternary <a,p,n>, where anchor *a* represents the selected object, positive *p* is the same category as anchor, while negative *n* is not the same category as anchor. The ultimate optimization target is to shorten the distance between *a* and *p*, and enlarge the distance between *a* and *n*. In the RMM module, the supervisory task of object detection uses Smooth L1 loss, which can control the gradient value of the outliers well:

$$Smooth\,L_1 = \begin{cases} 0.5x^2, & if\ |x| < 1 \\ |x| - 0.5, & otherwise. \end{cases} \quad (3)$$

Softmax Cross-Entropy is used as a classification loss in all SMA-Net modules. For an image, y represents the real ID of the image, and p_i represents the logit value of category i. The Softmax Cross-Entropy loss is performed as follows:

$$L_s = \sum_{i=1}^{N} -q_i log\,(p_i) \begin{cases} q_i = 0, & if\ y \neq i \\ q_i = 1, & if\ y = i . \end{cases} \quad (4)$$

In the training stage, we combine and optimize three kinds of losses to ensure that the PMA and RMM modules can converge simultaneously, as shown in the following formula:

$$T_1 = (L_{s1} + L_{S2} + L_{s3} + L_{s4} + L_{s5} + L_{s6} + L_{tri1} + L_{tri2} + L_{tri3}),$$

$$L_{total} = W_1 T_1 + (1 - W_1)\left(L_{s\_RMM} + L_{sl1}\right). \quad (5)$$

where $W_1 T_1$ represents the feature extraction loss in PMA. W_1 is a trade-off parameter and $L_{s\_RMM} + L_{sl1}$ represents the classification and regression loss in RMM. The network will rapidly converge by jointly training these two losses.

D. Pooling

In order to explore the impact of pooling on the PMA module, we test three different pooling methods into the PMA sub-network module, including GAP, GMP, and the GeM

pooling recently proposed. The calculation formula as follows:

$$e = \left[\left(\sum_{\mu\in\sigma} x_{cu}^{p}\right)^{\frac{1}{p}}\right]_{c=1\ldots C} \quad (6)$$

where $\mu \in \sigma = 1, \ldots, H \times 1, \ldots, W$ is a pixel on the feature map, c represents the channel of the feature map, x_{cu} is the corresponding tensor element, and p is a learnable parameter. When p is set to 1, the GeM pooling is equivalent to the GAP pooling. While $p = \infty$, the GeM Pooling is converted to the GMP pooling.

The response of the feature map will have a stronger localization ability when p increases. While training, GeM pooling aggregates features to an appropriately sized area by learning from the parameter p, which improves identification performance of the features. In the case of $p = 1$, GeM pooling is equivalent to the GAP pooling. In the case of $p = \infty$, the GeM Pooling is converted to the GMP pooling. The larger the p value, the more capable of locating the response of the feature map. During the training, GeM Pooling enables features to be more focused on an appropriately sized area by learning p value so as to improve the identification ability of the features.

IV. Dataset

To facilitate the study of Re-ID methods working on the same dressing site scene, we build a construction dataset named as GC-2245. This dataset is recorded from construction sites including the scenarios such as gate entrances, worksites, material processing sheds and so on. We divide the dataset into hard examples and easy examples according to the similarity degree of clothing. As shown in Fig. 3, person wearing reflective clothing and safety helmets are defined as hard examples, while others are defined as easy examples. There are 980 hard examples and 1265 easy examples in GC-2245.

Fig.3. Example images in our collected dataset GC-2245. On the left part some “easy” examples are shown, while other “hard” examples are shown on the right.

In the data collection process, night and irregular pose data are removed, the unsupervised clustering method is adopted for data pre-labelling, and precise identification is finally completed manually.

V. Experiments

A. Implementation details

Dataset. We have conducted comparative experiments on some of the most advanced person re-identification methods on the two datasets. 1) Person identification dataset GC-2245 collected from the construction site. This dataset contains 2245 IDs with a total of 54881 person images. Each ID contains 10-30 images, captured by at least 3 cameras. We use 1120 IDs for training, and the remaining 1125 IDs for model testing; 2) The public dataset Market1501 includes 1501 IDs with a total of 32668 person images captured by 6 cameras, 750 IDs of which are used for model training, and 751 IDs are used for testing.

The SMA-Net consists of two subnetworks: Object Detection (RMM) and Re-ID (PMA). Head and trunk coordinate information and person identification information will be used as supervisory labels during SMA-Net training. Therefore, both GC-2245 dataset and Market1501 dataset contain coordinates of the person's head and trunk (except occlusion), as well as identity information.

Training. The RMM needs to complete the supervision mapping of PMA in the SMA-Net network model. We divide the training process into two steps in order to make the network converge quickly, and select an SGD optimizer to optimize the network weights iteratively. We utilize the commonly used ResNet50 structure as the backbone network, and use the ImageNet pre-training model for parameter initialization. During the training, each mini-batch contain 16 IDs with a total of 64 person images, all of which have been adjusted to a uniform size before feeding to the network. We adopt several data enhancement methods in order to improve the robustness of the model, including horizontal flip, random erase and color dithering.

The RMM module is first trained as an object detection task, while the PMA sub-network is temporarily left out of training. The initial learning rate is set as 0.001, and 16 epochs are completed on the training dataset. The learning rate is updated to 0.1 times the original value at the 8-th and 14-th epochs, respectively. Considering that the detected target is the head and trunk, we remove the sub-branches conv9_2, conv10_2 and conv11_2 from the original SSD network used for finding large objects, and only keep the shallower sub-branch conv4_3, conv7, and conv8_2. In addition, we set anchor rate as {1, 2, 1/2} to avoid generating too many useless candidate boxes because the aspect ratios of the head and trunk targets are roughly in [1, 2].

The PMA sub-network continues to train after the RMM training is completed. As RMM and PMA share backbone network ResNet50, considering the great difference between RMM sub-network and PMA sub-network. The learning rate of PMA sub-network is set to 0.001 when it starts to train, while the learning rate of backbone network ResNet50 and RMM sub-network is set to 0.00001. During the training process, both the learning rate of the backbone network and the RMM sub-network will adopt the gradual warmup mechanism to grow to 0.001 within 10 epochs, and then will be decayed to 0.1 times. The whole training process will end after the 120-th epoch is completed.

Evaluation Metrics. We adopt the general indicator: mean average precision (mAP), to measure the identification

Fig.4. Examples of Re-ID results obtained by different methods. Red frames represent incorrect retrieval.

performance of SMA-Net and several current advanced Re-ID methods. All the experimental data are calculated by single query, other than post-processing methods such as re-ranking and multi-query fusion.

B. Comparison with the state of the art

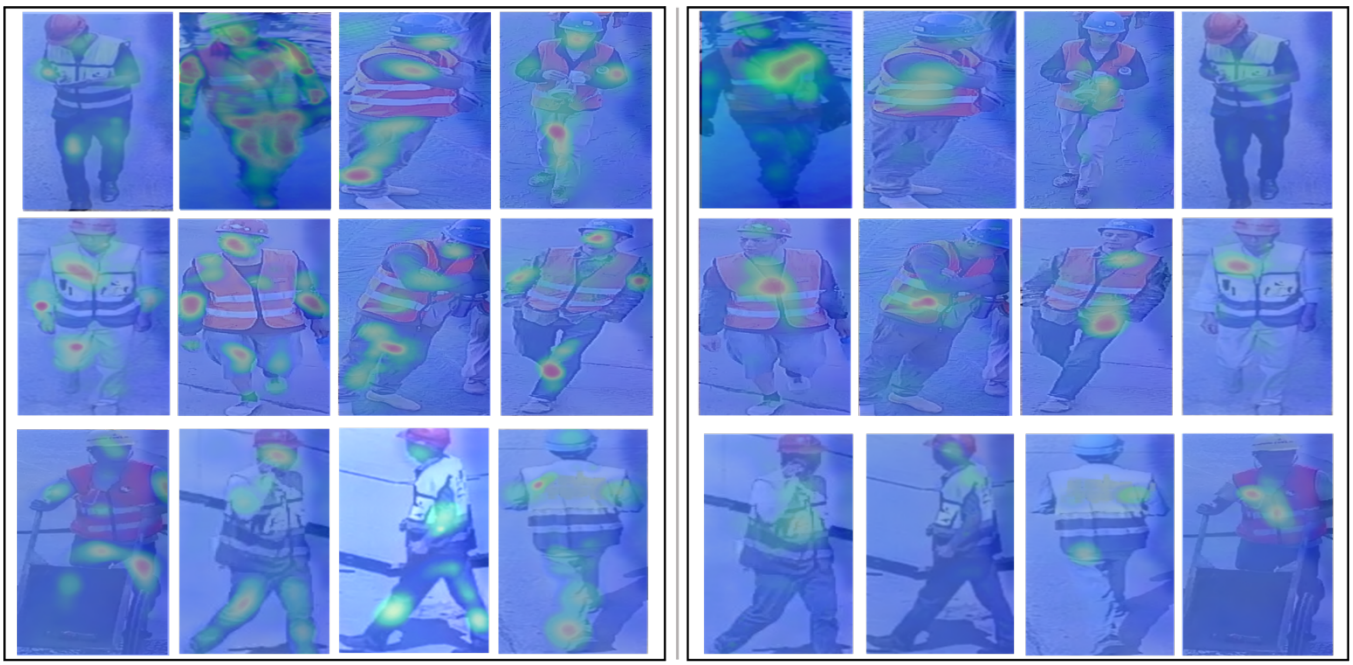
Results on GC-2245. We compare the performance of SMA-Net and current advanced Re-ID methods on dataset GC-2245 such as MGN [5], PCB + RPP [4], MaskReID [23], Mancs [24]. In Table I, it can be seen that SMA-Net has greatly improved the performance comparing to other Re-ID methods. SMA-Net represents ROI mask mapping method 1 and SMAS-Net represents ROI mask mapping method 2 in Fig. 2.

In the Hard set of the GC-2245 dataset, SMA-Net (with RMM method 1) achieves good results, with mAP and Rank-1 reaching 85.8% and 92.6%, respectively, 2.9% and 2.3% higher than the original MGN method. SMA-Net is increased by 6% and 4.5% in two Indicators mAP and Rank-1 compared with the original ResNet50 method. Meanwhile, SMA-Net also achieves the best results on the easy set, with mAP and Rank-1 reaching 87.5% and 94.6%, respectively, 3.2% and 2.8% higher than the original MGN method. SMA-Net is increased by 7.4% and 5.2% in mAP and Rank-1 compared with the original ResNet50 method. The above results show that our method can effectively solve the identification problem of extremely similar dress (e. g. uniform on the construction site).

TABLE I. PERFORMANCE COMPARISON BETWEEN SMA-NET AND SEVERAL CURRENT ADVANCED RE-ID METHODS ON THE DATASET GC-2245

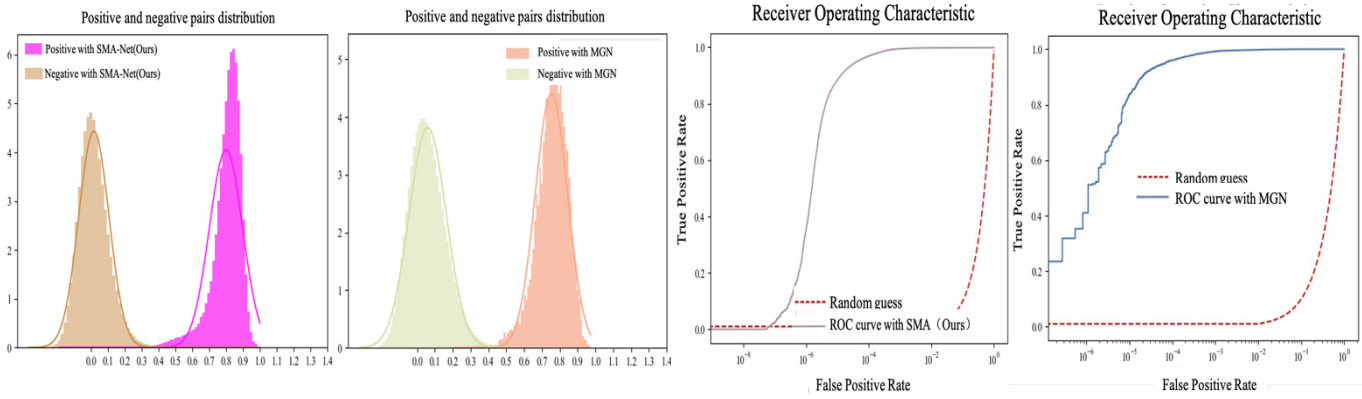
	GC-2245			
	Easy		Hard	
Method	mAP	Rank-1	mAP	Rank-1
ResNet50 [19]	80.4	89.4	79.8	88.1
MGN [5]	84.3	91.8	82.9	90.3
PCB-RPP [4]	82.2	91.1	81.3	89.4
MaskReID [23]	74.9	83.7	70.7	80.1
Mancs [24]	76.6	85.4	72.8	82.9
SMA-Net (ours)	87.5	94.6	85.8	92.6
SMAS-Net (ours)	**88.1**	**95.0**	**86.2**	**92.8**

SMAS-Net (with RMM method 2) achieved the best results, with a certain improvement in the mAP and Rank-1 compared with SMA-Net, but the complexity of network structure leads to an expensive computing cost and a slow inference speed. For this reason, we focus on the research on the low-cost SMA-Net.



(a) SMA-Net Heatmap (b) MGN Global and Partial Heatmap

Fig.5. Feature Attention Maps. (a) SMA-Net Partial Feature, (b) MGN Partial Feature.



(a) (b)

Fig.6. (a) Positive and negative sample pair distribution curve of SMA-Net and MGN (b) ROC curve of SMA-Net and MGN.

We compare the retrieval effects of SMA-Net and other methods on GC-2245. As some examples shown in Fig. 4, the person features extracted by MGN, PCB-RPP, and MaskReID will be significantly affected when the dress is extremely similar. The SMA-Net strengthen the attention on the features of the head and trunk parts, thus more identifiable person features being learned. The performance indicators of the above Re-ID model on the GC-2245 dataset are all lower than those of Market1501, which indicating that the performance of the Re-ID method is indeed degraded in the case of extremely similar dress.

As shown in Fig. 5, we visualize the feature heatmap of SMA-Net. The global feature representation and local feature representation of MGN [17] are on the right as a comparison. Although MGN [17] is a multi-granularity feature identification network, the feature attention of the network on body parts is not obvious due to the same clothing and clothing features. The SMA-net makes the network pay more attention to the local features. More attention is paid to the specific parts of local information suitable for the of datasets, in which the mosaics represent the feature mask.

As shown in Fig. 6 (a), the vertical coordinates represent the number of positive and negative sample pairs (unit K) and the horizontal coordinates represent their similarity. It can be observed that the discriminant ability of positive and negative samples of the SMA-Net is slightly stronger than that of MGN. At the same time, in Fig. 6 (b), the horizontal coordinates represent False Positive Rate (FPR) and the vertical coordinates represent represent True Positive Rate (TPR). When the FPR is the same, the TPR of SMA-Net is higher than that of MGN.

TABLE II. PERFORMANCE COMPARISON BETWEEN SMA-NET AND SEVERAL CURRENT ADVANCED RE-ID METHODS ON THE PUBLIC DATASET MARKET1501

	Market1501	
Method	mAP	Rank-1
ResNet50 [19]	84.6	93.3
SPReID [25]	81.3	92.5
Part-Aligned [8]	79.6	91.7
MaskReID [23]	75.3	90.0
AlignedReID [3]	79.3	91.8
PCB+RPP [4]	81.6	93.8
MGN [5]	86.9	**95.7**
Mancs [24]	82.3	93.1
DuATM [26]	76.6	91.4
HA-CNN [27]	75.7	91.2
SMA-Net (Ours)	**87.8**	**95.7**

Results on Market1501. We also compare the SMA-Net with the current state-of-the-art ReID methods on the Market1501 dataset. As can be seen from Table II, the SMA-Net method proposed by us has the best performance, with mAP and Rank-1 reaching 87.8% and 95.7%, respectively, 0.9% higher than the original MGN method.

VI. ABLATION STUDY

We conduct several ablation experiments on GC-2245 and Market1501 datasets to illustrate the effect of the pooling selection, RMM module and masking method on the performance of the SMA-Net.

TABLE III. THE IMPACT OF DIFFERENT POOLING METHODS IN THE PMA MODULE ON THE PERFORMANCE OF THE METHOD

	mAP	Rank-1
SMA-Net (GAP)	83.2	91.8
SMA-Net (GMP)	85.3	92.7
SMA-Net (GeM Pooling)	**86.6**	**93.8**

The Impact of GeM Pooling. We conducted three sets of comparative experiments on the dataset GC-2245, and the experimental results are shown in Table III. The results indicate that GMP pooling is better than GAP. GAP is pooling on the global feature map, which is easily interfered by background information and not differentiated in the attention on foreground information, while GMP overcomes this problem by aggregating the most differentiated features. Compared with GMP pooling, GeM Pooling achieves the best results, with 1.3% and 1.1% increases in mAP and Rank-1.

The Impact of RMM module. We conduct a series of comparative experiments to illustrate how RMM module works. The detailed results are shown in Table IV, which shows the performance of the two methods on the GC-2245 dataset, including all the hard and easy samples.SMA-Net (w/o RMM) represents the method that RMM module is removed by setting the loss weight *w1* in Equation (5) to 1 and adjusting the mask ratio in PMA to 0. As can be seen from the results in the table, the use of the RMM module on dataset GC-2245 can make performance indicators increased by 1.2% and 0.9% respectively on mAP and Rank-1, and increased by 0.5% and 0.3% respectively on dataset Market1501. This shows that the RMM module can significantly improve the identification performance, especially in the case of similar dressing.

The Impact of the masking method. In the SMA-Net, RMM employs the head and trunk coordinates of person to construct mask, and the PMA uses this mask template to results, and the results are shown in Table V. As can be seen from the table, when Mask-Ratio is 50%, the identification construct each sub-branch feature map. We conduct another set of comparative experiments in order to further explore the effect of the mask structure ratio on the final identification effects on both GC-2245 and Market15001 datasets. When Mask-Ratio is 0%, which is equivalent to removing RMM modules, the conclusion is in line with that in Table V. On the dataset GC-2245, the identification effects on mAP and Rank-1 are decreased by 6.4% and 5.3%, respectively, when Mask-Ratio is 100% (compared with 50%). On Market1501, the identification effects on mAP and Rank-1 are decreased by 4.4% and 2.9%, respectively, when Mask-Ratio is 100% (compared with 50%). Thus, discriminant information, such as body shape, silhouette and other identity information, still exists in similar dress areas.

TABLE IV. THE ABLATION EXPERIMENT RESULTS OF THE SUB-MODULE RMM OF THE SMA-NET ON THE DATASET GC-2245 AND THE PUBLIC DATASET MARKET1501

	GC-2245		Market1501	
Method	mAP	Rank-1	mAP	Rank-1
SMA-Net (w/o RMM)	85.4	92.9	87.3	95.4
SMA-Net (RMM)	**86.6**	**93.8**	**87.8**	**95.7**

TABLE V. THE ABLATION EXPERIMENT RESULTS OF THE SUB-MODULE RMM OF THE SMA-NET ON THE DATASET GC-2245 AND THE PUBLIC DATASET MARKET1501

		GC-2245		Market1501	
Method	Mask-Ratio	mAP	Rank-1	mAP	Rank-1
SMA-Net	0	85.4	92.9	87.3	95.4
	0.5	**86.6**	**93.8**	**87.8**	**95.7**
	1	80.2	88.3	83.4	92.8

VII. Conclusion

For the problem of Re-ID in similar dressing scenarios, we propose an multi-task person re-identification network SMA-Net, and also build an person re-identification dataset GC-224 for site scenarios. The RMM module provide supervisory mapping for ROI parts, and the PMA combine attention modules in the local feature learning framework, which improve the model's feature extraction ability for the same dress accessories. Besides, we demonstrate the effectiveness of SMA-Net in solving similar dressing problems. On the GC-2245 dataset, 6.65% of mAP is improved, the accuracy of Rank-1 is improved by 3.8%, and the optimal performance is obtained when compared with other methods.

References

[1] L. Zheng, L.Y. Shen, L. Tian, S.J. Wang, J. D. Wang, and Q. Tian, " Scalable Person Re-identification: A Benchmark." IEEE International Conference on Computer Vision, 2015, pp.1116-1124.

[2] H.Y. Zhao, M.Q. Tian, S.Y. Sun, J. Shao, J.J. Yan, S. Yi, X.G. Wang, and X.O. Tang, "Spindle Net: Person Re-identification with Human Body Region Guided Feature Decomposition and Fusion." IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp.907–915.

[3] H. Luo, W. W. Jiang, X.Zhang, X. Fan, J.J. Qian, and C. Zhang, "AlignedReID++: Dynamically matching local information for person reidentification". Pattern Recognit. 94, 2019, pp.53–61.

[4] Y.F. Sun, L. Zheng, Y. Yang, Q. Tian, and S.J. Wang, "Beyond Part Models: Person Retrieval with Refined Part Pooling". European Conference on Computer Vision, 2018, pp.480-496.

[5] G.S. Wang, Y.F. Yuan, X. Chen, J.W. Li,and X. Zhou, "Learning Discriminative Features withMultiple Granularitiesfor PersonRe-Identification". ACM international conference on Multimedia, 2018, pp.274-282.

[6] J. Xu, R. Zhao, F. Zhu, H.M. Wang, and W.L. Ouyang, "Attention-aware compositional network for person re-identification". IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp.2119–2128.

[7] C.F. Song, Y. Huang, W.L. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification". IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp.1179-1188.

[8] Y.M. Suh, J.D. Wang, S.Y. Tang, T. Mei, and Kyoung Mu Lee, "Part-aligned bilinear representations for person re-identification". European Conference on Computer Vision, 2018, pp.402-419.

[9] T. Chen, et al. "ABD-net: Attentive but diverse person re-identification." IEEE International Conference on Computer Vision, 2019, pp.8351-8361.

[10] Z. Zhang, C. Lan, W. Zeng, et al, "Relation-Aware Global Attention for Person Re-Identification" IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2020, pp.3186-3195.

[11] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems 27, 2014.

[12] Y. Latif, et al. "Addressing Challenging Place Recognition Tasks using Generative Adversarial Networks." IEEE International Conference on Robotics and Automation, 2018, pp.2349-2355. .

[13] Y.X. Ge, Z.W. Li, H.Y. Zhao, G.J. Yin, X.G. Wang, and H.S. Li, "FD-GAN: Pose-guided feature distilling GAN for robust person re-identification." arXiv:1810.02936, 2018.

[14] Y. Huang, J.S. Xu, Q. Wu, Z.D. Zheng, Z.X. Zhang, and J. Zhang, "Multi-pseudo regularized label for generated samples in person re-identification." IEEE Transactions on Image Processing, 2018, pp.1391-1403.

[15] J.X. Liu, B.B. Ni, Y.C. Yan, P. Zhou, S. Cheng, and J.G. Hu, "Pose transferrable person re-identification". IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp.4099-4108.

[16] X.L. Qian, Y.W. Fu, T. Xiang, W.X. Wang, J. Qiu, Y. Wu, "Pose-normalized image generation for person re-identification". European Conference on Computer Vision, 2018, pp.650-667.

[17] Z. Zheng , X. Yang , Z. Yu , et al, "Joint Discriminative and Generative Learning for Person Re-Identification". IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp.2138-2147.

[18] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and F.F. Li. 2009, "ImageNet: A large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition", 2009, pp.248–255.

[19] K.M. He, X.Y. Zhang, S.Q. Ren, and J. Sun, "Deep Residual Learning for Image Recognition." IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp.770–778.

[20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, and Alexander C Berg, "SSD: Single shot multibox detector." European Conference on Computer Vision, 2016, pp.21–37.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation". International Conference on Medical image computing and computer-assisted intervention, 2015, pp.234–241.

[22] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification." arXiv:1703.07737, 2017.

[23] L. Qi, J. Huo, L. Wang, Y.H. Shi, and Y. Gao. "Maskreid: A mask based deep ranking neural network for person re-identification." arXiv preprint arXiv:1804.03864, 2018.

[24] C. Wang, Q. Zhang, C. Huang, W.Y. Liu, and X.G. Wang, "MANCS: A Multi-task Attentional Network with Curriculum Sampling for Person Re-Identification". European Conference on Computer Vision, 2018, pp.365-381.

[25] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification." IEEE conference on computer vision and pattern recognition 2018, pp.1062-1071.

[26] J. Si, H. Zhang, C. G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang, "Dual attention matching network for context-aware feature sequence based person re-identification."IEEE conference on computer vision and pattern recognition, 2018, pp.5363-5372.

[27] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification." IEEE conference on computer vision and pattern recognition, 2018, pp.2285-2294.