Weak Reverse attention with Context Aware for Person Re-identification

Ke Gong [2,3,*,a], Xin Ning [1,2,3,*,b], Hanchao Yu [4], Liping Zhang[1,2,3] and Linjun Sun [1,2,3]

[1] Laboratory of Artificial Neural Networks and High-speed Circuits, Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China
[2] Cognitive Computing Technology Joint Laboratory, Wave Group, Beijing 100083
[3] Beijing Wave Security Technology company lmited, Beijing, 102208
[4] Bureau of Frontier Sciences and Education, Chinese Academy of Sciences, Beijing, 100864

Corresponding Author: [a]gongke@wavewisdom-bj.com; [b]ningxin@semi.ac.cn;

Abstract. Person re-identification is a difficult topic in computer vision. Some study think that current deep learning methods is biased to capture the most discriminative features and ignore low-level details, more serious is it pay too much attention on relevance between background appearances of person images. It might limit their accuracy or makes them needlessly expensive for a not best performance. In this paper, we carefully design the Weak Reverse attention with Context Aware Network (WRCANet). Specifically, by merging weak reverse attention network and content aware module, the model can not only remove the background noise to extract the main information of persons, but also suppress the loss of local detailed information as the network deepens. We experiment on the Market-1501, DukeMTMC-reID and CUHK03, and the results show that our method achieves the state-of-the-art performance.

1. Introduction

Person re-ID has received more and more attention from both the scientific and industrial community in recent years due to its wide application prospects. At present, deep learning method has improved the performance of person re-ID to a new level [1-4]. However, some challenges remain.

Firstly, the existing methods almost ignore the impact of the background region on the performance of the person re-ID model. In the field of re-ID, background information has a significant impact on the performance of models[5]. Secondly, in the task of image recognition, more attention is paid to extract the most discriminative features, so with the deepening of network depth, more and more details are ignored.

To solve these problems, we propose Weak Reverse attention with Context Aware network (WRCAnet) for person re-ID. Compared with previous papers, there are the following three contributions in this paper:

- We propose a weak anti-attention mechanism to eliminate background-bias for Robust Person Re-identification. It should be the first time that weak reverse attention has been proposed.
- We designed a content perception module based on multi-stage feature fusion and multi-granularity feature fusion, which can extract multi-scale and more refined features.
- The experimental results show that the proposed method achieves state-of-the-art performance

on the three benchmarks of Market-1501, DukeMTMC-reID and CHUK03.

2. Related Work

In recent years, attention mechanisms have been widely used in re-ID[6-8]. [6] uses a combination of spatial attention and channel attention to extract global features, and uses channel attention to extract local features. Li et al. [7] adds a multi-branch harmonious attention mechanism to Resnet to reduce the number of parameters and extract multi-local features. In addition, attention mechanism based on attribute guidance and attitude estimation has been widely used. Xu J et al.[8]utilized pose-guided part Attention to locate body part so that can Extract pedestrian features to re-ID.

It has been proved that using multi-level method is a powerful technique for learning high-level semantic features and solving the issue of misalignment of body parts [9-12]. In this paper, a content aware module is designed, which is composed of multi-scale feature fusion and multi-granularity feature fusion to integrate more useful information to prevent information loss with the deepening of the network.

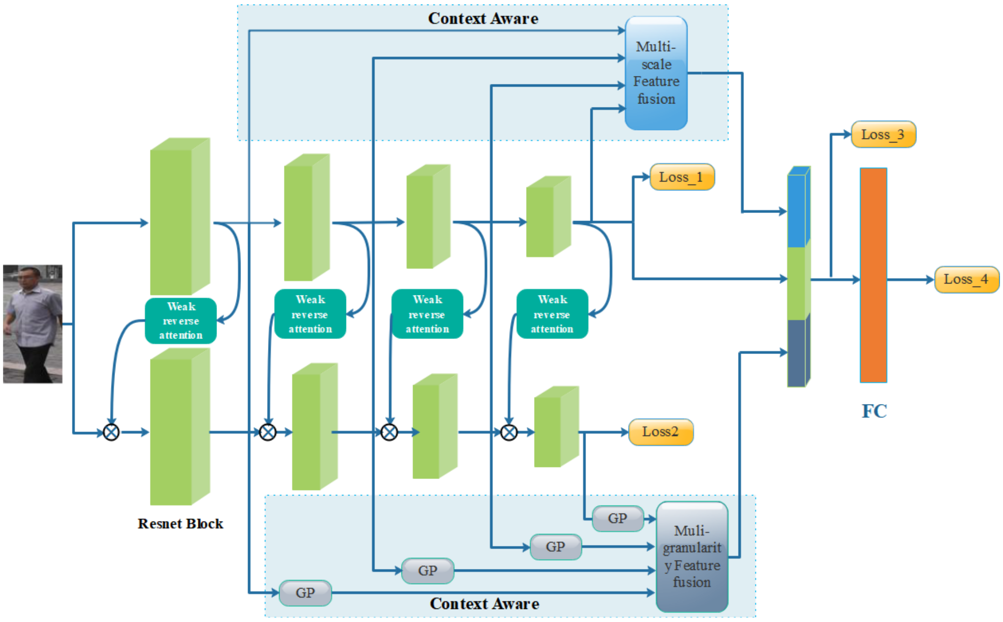

Figure 1. Framework of proposed Weak Reverse attention with Context Aware network (WRCAnet) for person re-ID. It contains backbone network (ResNet-50+ "Weak Reverse attention") and Contest Aware subnetwork. The Contest Aware subnetwork is composed of multi-scale branches and multi granularity branches.

3. Proposed Method

3.1. Overall Architecture

In this paper, shown in Figure 1, we build the model based ResNet-50. In order to obtain larger feature map and more information, we modified the stride of last down-sampling operation from 2 to 1, do as the successful methods[1] do.

According to [13], the Resnet consists of 4 residual modules. The output features are also different at different stages. In the lower stage, feature maps often have good spatial position expression capabilities. While in the high stage, they have strong contextual ability and more semantic information. However, no matter in what stage, it extracts the global information of the image, which has the ability to express the global information. Multi-level fusion can complement the defects of deep features and shallow features. Therefore, we design a channel attention method similar to Senet [14] to fuse features as an expression of multi-scale information.

3.2. Backbone Network with Weak Reverse Attention

We propose a model Resnet + weak reverse attention model different from the existing methods[8], as shown in Figure 2.

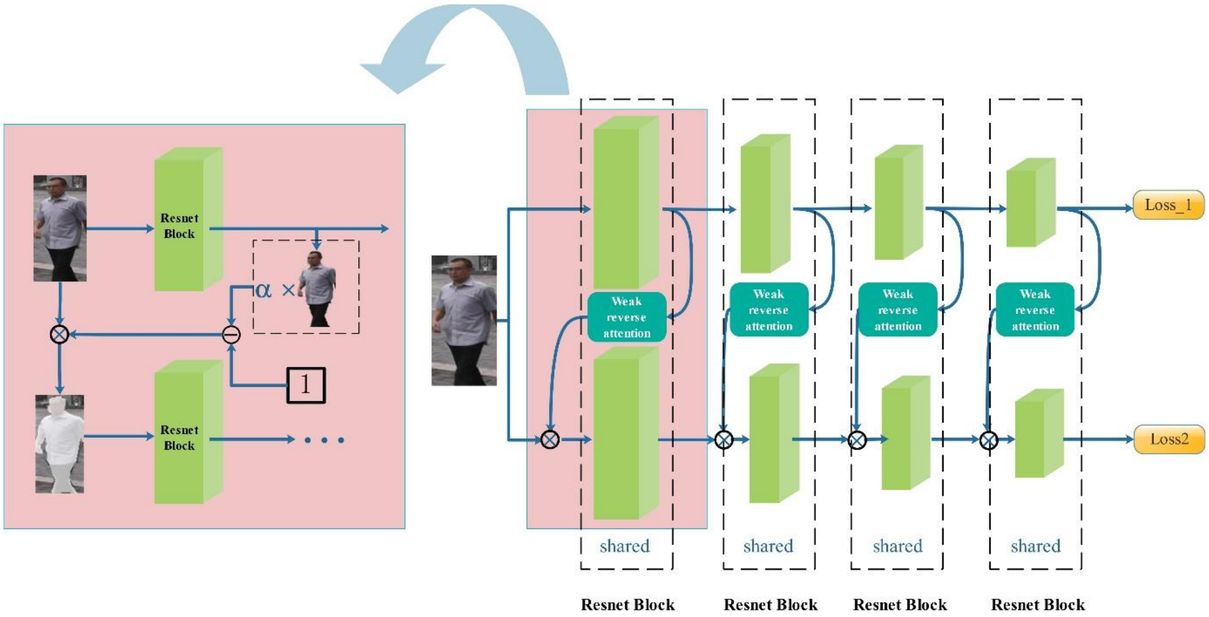

Figure 2. Framework of the backbone network.

Although there are two Resnet streams in the figure, they share parameters, so the parameters are not increased. The left is the decomposition of the WRA. As shown in Figure 2, the two Resnet streams share the same parameters, which effectively prevents the model from overfitting. We remove the most interested areas of the network in a certain proportion, not completely, so that the network can extract the original attention area, also pay more attention to extract other useful features.

The process of weak reverse attention is shown on the left in Figure 2. Specifically, for a feature $A_{ij} \in R^{C\times W\times H}$ which is after the residual module, we define the weight of each channel α_c as:

$$\alpha_c = \frac{1}{H\times W}\sum_i\sum_j A_{ij} \tag{1}$$

in order to get the attention map, the channel needs to be reduced to a single channel, so the weighted average operation is done along the channel. We define the upsampling operation $up\_sampe(\cdot)$ to align the size of the feature map with the input. After the activation function $relu$, the most discriminative feature map f_{ij} described above is defined as formula (2):

$$f_{ij} = \mathrm{Re}\,lu(up\_sampe(\frac{1}{C}\sum\alpha_c \otimes A_{ij}) \tag{2}$$

The input feature minus this most discriminative feature map, which is the same size as the input feature, is the reverse attention described in [1]. We define weak reverse attention as formula (3). The parameter $\alpha \in (0,1)$ is the weakening coefficient, which is 0.5 in this paper.

$$f_{new} = I \otimes (1-\alpha\cdot sigmoid(f_{ij})) \tag{3}$$

3.3. Context Aware

Content perception in this paper is composed of multi-scale fusion module and multi-granularity fusion module.

Multi-scale Feature fusion. In different stages, the residual module outputs different sizes of feature maps. This paper compressed the output feature of the four modules of Resnet to the same size

(24*12) through the AdaptiveAvgPool operation. Then, the weight of each group of features is obtained through global pooling, and a 1-dimensional convolution is used to fuse the four feature graphs to get a 1* 288-dimensional fusion feature, as shown in figure 3.

Multi-granularity Feature fusion. In order to obtain the multi-granularity features of the image, we designed a GP branch network, as shown in Figure 4. The global maximum pooling (GMP) feature and global minimum pooling (GAP) feature are taken for the output of each level of residual module, then the sum of the pooling features is calculated, and then a convolution layer whose kernel size and stride is 1 is connected to reduce the channels to 256. Figure 5 shows the structure of a multi-granularity feature fusion network. Similar to multi-scale feature fusion, we connect four 1 * 256 features and then use a one-dimensional convolution to fuse to a 1 * 256-dimensional multi-granularity feature.

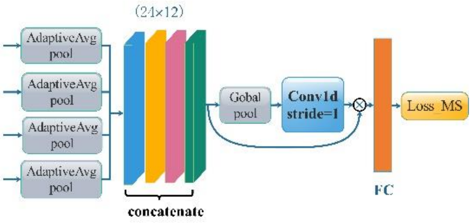

Figure 3. Multi-scale Feature fusion. The channel attention is achieved by global pooling and one-dimensional convolution.

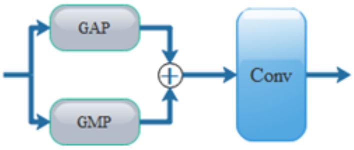

Figure 4. GP net. It is composed of maximum pooling, average pooling and convolution with kernel=1.

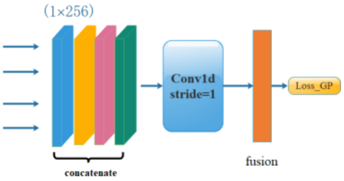

Figure 5. Multi-granularity Feature fusion. The input features are connected, and convolution is used to fuse these features.

3.4. Loss function

In the network, we use the sum of loss functions in each phase of the network as the final loss function:

$$L_{all} = Loss\_1 + Loss\_2 + Loss\_3 + Loss\_4 + Loss\_GP + Loss\_MS \tag{4}$$

Different loss functions are used in different stages, among which , use the softmax loss with label smoothing regularization[15], and the rest use the batch hard triplet loss[16]. By calculating the weighted average and average distribution of the hard target in the dataset, the softmax loss with label smoothing regularization can effectively reduce overfitting on small sample data sets:

$$L_{sof-LS} = -\frac{1}{N}\sum_{i=1}^{N} g_i \log((1-\varepsilon)p_i + \frac{\varepsilon}{S} \tag{5}$$

The traditional triples loss[17] randomly sample three pictures from the training data. For mining hard samples pairs, the batch hard triplet loss is proposed. And it enhances the compactness of intra-class and the separability of inter-class, which is defined as:

$$L_{tri} = \sum_{i=0}^{P}\sum_{a=1}^{K}[m + \max_{p=1\ldots K} D(f(I_a^i), f(I_p^i)) - \min_{\substack{j=1\ldots P\\ n=1\ldots K\\ j\neq a}} D(f(I_a^i), f(I_n^i))]_+ \tag{6}$$

4. Experiment and results

4.1. Result

In order to prove the superior performance of our method, we compare it with the best results at present. The specific results are as follows:

Table 1. Comparison of WRCAnet and other state-of-the-art methods on the Market-1501, DukeMTMC-ReID and CUHK03(Labelled and Detected) datasets. The bold are the best result.

Method	Market-1501		DukeMTMC-reID		CUHK03(L)		CUHK03(D)	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP

HA-CNN[7]	91.2%	75.7%	80.5%	63.8%	44.4%	41.0%	41.7%	38.6%
DaRe[10]	90.9%	86.7%	84.4%	80.4%	73.8%	74.4%	84.4%	80.0%
Mancs[18]	93.1%	82.3%	84.9%	71.8%	69.0%	63.9%	65.5%	60.5%
MGN[19]	95.7%	86.9%	88.7%	78.4%	66.8%	66.0%	68.0%	67.4%
MGCAN[20]	94.7	87.1	-	-	50.14	50.21	46.71	46.87
BDB+Cut[21]	95.3%	86.7%	89.0%	76.0%	79.4%	76.7%	76.4%	73.5%
Ours	**96.1%**	**87.8%**	87.5**%**	**79.6%**	**81.6%**	**78.7%**	**81.4%**	**78.9%**

Market-150, from table 1, our model achieves relatively advanced performance mAP/Rank-1=87.8%/96.1%. By applying re-ranking[22], we can further get the better result that mAP and rank-1 are improved by 1.3% and 4.4%.

On DukeMTMC-reID, it can be see that although our results are not the best, our model achieves mAP/Rank-1=79.6%/87.5%. By applying re-ranking, we can further get the better result that mAP and rank-1 are improved by 5.3% and 3.8%.

On CHUK03, the result of our model is much higher than that of other models on labeled (map / rank-1 = 78.7% /81.6 %); on the detected dataset, we also get the result of map / rank-1 =78.9% /81.4 %.

4.2. Ablation study

As mentioned above, our method mainly consists of three parts: a) weak reverse attention; b) Multi-granularity Feature fusion module; and c) Multi-scale Feature fusion module. To evaluate the impact of each part on the experiment, we also performed additional experiments. we conducted the ablation study of model components on market-1501 dataset to analyze the influence of the above components on the model.

Table 2. The result of ablation study on Market-1501. WRA represents weak reverse attention. GP represents GP branches which include GP net and Multi-granularity Feature fusion net. MS represents Multi-scale Feature fusion net.

Model	mAP	Rank-1	Rank-5	Rank-10
ResNet-50	71.4%	87.5%	94.9%	96.7%
ResNet-50+ WRA	78.9%	90.2%	95.9%	97.3%
ResNet-50+ WRA+GP	83.3%	92.8%	97.0%	98.2%
ResNet-50+ WRA+MS	86.2%	93.5%	97.7%	98.3%
ResNet-50+ WRA+MS+GP	87.8%	96.1%	98.5%	99.1%

The impact of WRA. Apply weak reverse attention (WRA) to resnet-50 as our backbone network. Comparing with resnet-50, we can find that without introducing more parameters than resnet-50, the mAP and Rank-1 of the ResNet-50+ WRA in this paper improved by 7.5% and 2.7%, respectively.

The impact of Multi-granularity Feature fusion. In the experiment, we added multi-granularity fusion branch to the main network and removed multi-granularity branch from the whole network. From the results in table 2, we found that the network could improve from mAP/Rank-1=78.9%/90.2% to mAP/Rank-1=83.3%/92.8% after adding multi-granularity branches. And mAP and Rank-1 declined by 1.6% and 1.8%, respectively, after removing the multi-grained branches.

The impact of Multi-scale Feature fusion. As shown in table 2, "resnet-50 + WRA+MS" is 7.3% and 2.8% higher than "resnet-50 + WRA". Simultaneously, As shown in table 2, "resnet-50 + WRA+MS" improved by 7.3% and 2.6% compared with "resnet-50 + WRA", which decreased much more than the multi-granularity fusion module. It also shows to a certain extent that multi-scale fusion in this paper contributes more to the model than multi-granularity fusion.

5. Conclusion

For this paper, Our aim is to introduce a new deep learning approach for person re-ID. We apply weak reverse attention mechanism, combined with content aware module which is composed of multi-granularity and multi-scale to learn the global and local features of the picture without introducing image segmentation, attribute recognition and pose estimation, to achieve an end-to-end person re-recognition model. We conducted part estimation experiments on the market dataset, and conducted a large number of experiments on three mainstream public datasets, and the results proved that our method have achieved state-of-the-art results.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 61901436).

References

[1] Sun Y, Zheng L, Yang Y, Tian Q and Wang S 2017 Beyond Part Models: Person Retrieval with Refined Part Pooling *Eur. Conf. Comput. Vis.* 1–17

[2] He L, Liang J, Li H and Sun Z 2018 Deep Spatial Feature Reconstruction for Partial Person Re-identification: Alignment-free Approach *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2** 7073–82

[3] Si J, Zhang H, Li C G, Kuen J, Kong X, Kot A C and Wang G 2018 Dual Attention Matching Network for Context-Aware Feature Sequence Based Person Re-identification *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 5363–72

[4] Kalayeh M M, Basaran E, Gokmen M, Kamasak M E and Shah M 2018 Human Semantic Parsing for Person Re-identification *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 1062–71

[5] Tian M, Yi S, Li H, Li S, Zhang X, Shi J, Yan J and Wang X 2018 Eliminating Background-bias for Robust Person Re-identification *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 5794–803

[6] Zhu Y, Guo X, Liu J and Jiang Z 2019 MULTI-BRANCH CONTEXT-AWARE NETWORK FOR PERSON RE-IDENTIFICATION Beijing University of Posts and Telecommunications , Beijing , China Academy of Broadcasting Science , Beijing , China *2019 IEEE Int. Conf. Image Process.* 2274–8

[7] Li W, Zhu X and Gong S 2018 Harmonious Attention Network for Person Re-identification *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 2285–94

[8] Xu J, Zhao R, Zhu F, Wang H and Ouyang W 2018 Attention-Aware Compositional Network for Person Re-identification *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 2119–28

[9] Luo J, Liu Y, Gao C and Sang N 2019 LEARNING WHAT AND WHERE FROM ATTRIBUTES TO IMPROVE PERSON Jinghao Luo , Yaohua Liu , Changxin Gao ∗ , Nong Sang Key Laboratory of Ministry of Education for Image Processing and Intelligent Control , School of Artificial Intelligence and Automation , Huazh *2019 IEEE Int. Conf. Image Process.* 165–9

[10] Wang Y, Wang L, You Y, Zou X, Chen V, Li S, Huang G, Hariharan B and Weinberger K Q 2018 Resource Aware Person Re-identification Across Multiple Resolutions *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 8042–51

[11] Lan X, Zhu X and Gong S 2018 Person Search by Multi-Scale Matching *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **11205 LNCS** 553–69

[12] Guo Y and Cheung N M 2018 Efficient and Deep Person Re-identification Using Multi-level Similarity *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **1** 2335–44

[13] Yu C, Wang J, Peng C, Gao C, Yu G and Sang N 2018 Learning a Discriminative Feature Network for Semantic Segmentation *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*

[14] Hu J, Shen L and Sun G 2018 Squeeze-and-Excitation Networks *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*

[15] Szegedy C, Vanhoucke V, Ioffe S, Shlens J and Wojna Z 2016 Rethinking the Inception Architecture for Computer Vision *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*

[16] Hermans A, Beyer L and Leibe B 2017 In Defense of the Triplet Loss for Person Re-Identification

[17] Schroff F, Kalenichenko D and Philbin J 2015 FaceNet: A unified embedding for face recognition and clustering *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*

[18] Zhong Z, Zheng L, Zheng Z, Li S and Yang Y 2018 Camera Style Adaptation for Person Re-identification *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 5157–66

[19] Li D, Chen X, Zhang Z and Huang K 2017 Learning deep context-Aware features over body and latent parts for person re-identification *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*
[20] Wang G, Yuan Y, Chen X, Li J and Zhou X 2018 Learning discriminative features with multiple granularities for person re-identification *MM 2018 - Proc. 2018 ACM Multimed. Conf.* 274–82
[21] Song C, Huang Y, Ouyang W and Wang L 2018 Mask-Guided Contrastive Attention Model for Person Re-identification *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 1179–88
[22] Zhong Z, Zheng L, Cao D and Li S 2017 Re-ranking person re-identification with k-reciprocal encoding *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*