Supplementary Material of MSINet: Twins Contrastive Search of Multi-Scale Interaction for Object ReID

Jianyang Gu[1] Kai Wang[2] Hao Luo[3] Chen Chen[4] Wei Jiang[1*]
Yuqiang Fang[5] Shanghang Zhang[6] Yang You[2] Jian Zhao[7*]
[1]Zhejiang University [2]National University of Singapore [3]Alibaba Group
[4]OPPO Research Institute [5]Space Engineering University [6]Peking University
[7]Institute of North Electronic Equipment
{gu_jianyang, jiangwei_zju}@zju.edu.cn zhaojian90@u.nus.edu

1. Search on VeRi-776

We select a training-validation ratio of 60%-80% in the searching process on MSMT17 [6]. Without changing any specific configurations, we directly search for the rational interaction operations on VeRi-776 [3, 4] dataset. The searched architecture is denoted as MSINet-VR. We compare the structure of MSINet and MSINet-VR in Tab 7. Generally, the two searched architecture have common characteristics: Channel Gate is preferred in shallow layers, while Cross Attention is employed for more thorough information interaction in deep layers.

Quantitatively, we also conduct relevant supervision and cross-domain experiments with MSINet-VR in Tab. 8. All the experiment configurations are kept the same as those of MSINet training. Although there are some fluctuations, generally MSINet-VR has similar performance to MSINet, and the retrieval accuracy still surpasses ResNet50 [1, 5] by a large margin.

2. Search with Different Overlap Ratios

With the identities of training and validation sets unbound, we conduct a series of experiments utilizing different data separation ratios in Tab. 9 to find the appropriate interaction operations for the network. Firstly, we separate the training and validation sets completely with no identity overlaps. It can be observed that a balanced train-validation ratio generally brings better performance. For the two extremes of data distribution, an over-small validation set makes the architecture optimizer stuck in local minima and achieves poor performance. On the contrary, an over-large validation set brings no severe damage to the architecture search process, despite that the model is still not optimal. It demonstrates that abundant validation data is essential for ReID NAS.

*Co-corresponding authors

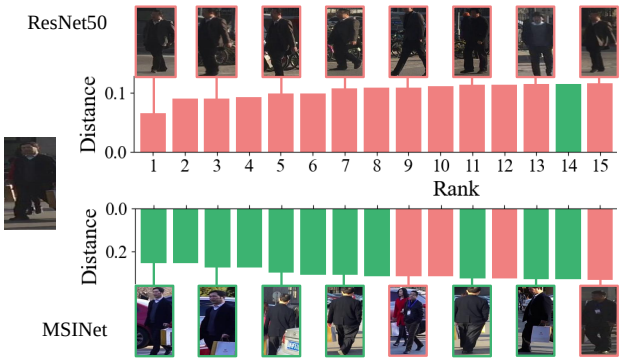

Figure 1. Example top-20 retrieved sequences comparison on MSMT17. Best viewed in color.

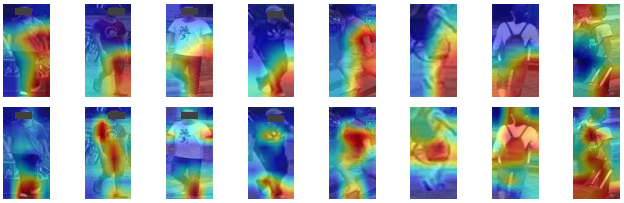

Figure 2. Example activation maps of ResNet50 (the first row) and our proposed MSINet (the second row) trained on Market-1501 dataset. Best viewed in color.

Secondly, we randomly select part of identities, and evenly divide their images into the training and validation sets. The experiment results suggest that having a relatively small proportion of overlapped identities, whose images have been partly utilized for model parameter update, stabilizes the searching process and leads to a better architecture. However, when the overlap increases to a certain extent, the resemblance between the training and validation sets will bring negative influence to the ReID architecture search. As a comparison, we conduct the search task with

Table 7. The detailed interaction operation comparison between MSINet, MSINet-VR and MSINet-S. N: None; E: Exchange; G: Channel Gate; A: Cross Attention.

Model	Cell #1		Cell #2		Cell #3		Cell #4		Cell #5		Cell #6	
MSINet	1	2	3	4	5	6	7	8	9	10	11	12
	G	G	E	G	A	G	G	N	G	A	E	A
MSINet-VR	1	2	3	4	5	6	7	8	9	10	11	12
	G	G	E	A	G	A	G	A	A	A	E	A
MSINet-S	1	2	3	4	5	6	7	8	9	10	11	12
	E	A	G	A	A	A	G	A	E	A	E	A

Table 8. Supervised performance on object ReID datsets. The results in the top part are trained from scratch, and those in the bottom part are pre-trained on ImageNet in advance.

Method	Params	Inference Time	M		MS		VR		VID		MS→M		VR→VID	
			R-1↑	mAP↑	R-1↑	mAP↑	R-1↑	mAP↑	R-1↑	R-5↑	R-1↑	mAP↑	R-1↑	R-5↑
ResNet50* [5] ∼	24M	1x	85.7	68.3	48.0	25.7	92.8	69.9	70.6	76.6	-	-		
MSINet	2.3M	0.71x	94.6	87.0	**76.0**	**52.5**	**95.9**	75.0	**76.5**	**89.8**	-	-	-	-
MSINet-VR	2.3M	0.71x	**94.7**	**87.4**	75.5	51.8	**95.9**	**76.0**	75.2	86.3	-	-	-	-
ResNet50* [5]	∼24M	1x	94.5	85.9	75.5	50.4	94.5	73.6	76.5	89.9	58.8	31.8	42.8	61.9
MSINet	2.3M	0.71x	95.3	**89.6**	**81.0**	**59.6**	96.8	**78.8**	**77.9**	**91.7**	**74.9**	**46.2**	48.0	65.6
MSINet-VR	2.3M	0.71x	**95.4**	89.0	80.1	57.5	**97.0**	78.6	77.3	91.3	72.9	44.8	**48.5**	**66.2**

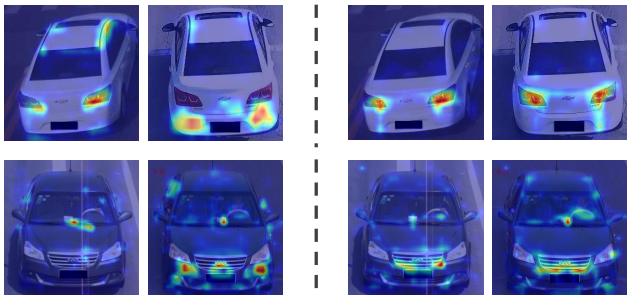

Figure 3. Example activation maps of MSINet trained on the task of VR→VID. With the Spatial Alignment Module, the model is capable to focus consistently on specific areas confronted with images from different sources. Best viewed in color.

traditional NAS scheme where a linear classification layer and cross entropy loss are employed for the training and validation data, the searched model of which performs worse than our proposed TCM.

Combined with above rules and the model performance, we select the architecture searched with the train-validation split of 60%-80% as the proposed MSINet.

3. Search with Softmax Loss

We further compare the detailed interaction operations between MSINet and the architecture searched under traditional NAS scheme, where softmax loss and a unified linear classification layer are utilized for the training and validation sets [2] (denoted as MSINet-S) in Tab. 7. Compared with MSINet and MSINet-VR, where direct information exchange mainly appears at deep layers, MSINet-S contains a large amount of Exchange and Cross Attention along the whole network. The over-frequent information exchange fails to focus on discriminative features. It also validates the effectiveness of our proposed Twins Contrastive Mechanism on searching for architectures suitable for ReID.

4. Visualization Results

Some additional visualization results are illustrated to further manifest the effectiveness of our proposed architecture. Firstly, we visualize an example comparison of the top-20 retrieved sequences between ResNet50 and MSINet on MSMT17 in Fig. 1. ResNet50 mainly focus on general appearance information, while our proposed MSINet concentrates on discriminative distinctions, the hand bag in this case. Even though positive samples have large appearance differences from the query image, MSINet is still capable to distinguish them.

Secondly, example activation maps of ResNet50 and our proposed MSINet on Market-1501 [7] are visualized in Fig. 2. ResNet50 mainly focuses on the right part of the image, including some background areas. Our proposed MSINet, oppositely, is capable to dynamically focus on discriminative distinctions of each image.

Thirdly, to intuitively demonstrate the effectiveness of the Spatial Alignment Module (SAM) on enhancing the at-

Table 9. The evaluation results of models searched with different training-validation identity ratios on the MS dataset. * indicates that the model is searched with the softmax loss.

w/o overlap				w/ overlap			
train (%)	valid (%)	MS R-1↑	MS mAP↑	train (%)	valid (%)	MS R-1↑	MS mAP↑
90	10	70.6	46.2	60	60	75.3	51.2
75	25	71.4	46.9	40	80	75.1	50.5
67	33	75.5	51.5	**60**	**80**	**76.0**	**52.5**
50	50	74.7	50.4	80	60	74.8	51.0
33	67	74.7	50.8	80	80	74.4	50.3
25	75	74.9	50.6	100	100	74.4	50.1
10	90	75.4	50.7	100*	100*	74.0	49.8

tention consistency of the model confronted with images from different sources, we visualize example activation maps on the task of VR→VID. As shown in Fig. 3, without alignment, the model can have different activated positions on different images of the same identity, even if they share similar appearances.

5. Comparison and advantages to OSNet

(1) OSNet simply sums up the features of each branch, without detailed exploration on the interaction between branches. In comparison, MSINet practically select rational interaction operations for different network layers. Consequently, MSINet surpasses OSNet not only in supervised, but also in domain generalization performance by a large margin. (2) OSNet contains 4 branches with different receptive field scales, where there exists certain parameter redundancy. We validated in the early exploring that removing the branches with receptive field scales of 3 and 5 has little influence to the model performance. MSINet reduces the number of branches, and increases the scale difference between two branches, which increases the parameter amount by a little bit but significantly reduces the inference time.

6. Detailed Analysis on SAM

We compare the proposed SAM module to some previous attention-based methods and analyze it in detail. [9] regularizes the attention generated at different network layers for the same image; [8] explicitly enforces the longitudinal activation distribution to be the same for two images, which may lead to misalignment if the objects are not properly detected. For negative samples, there can be many different hints for recognition, some of which might be inappropriate, such as the backgrounds. By aggregating the information from different negative samples, the network is driven to only focus on discriminative regions. The motivation of SAM is different from the above two works. For the in-domain setting, the camera condition diffrences are directly addressed by supervised learning. Thus, SAM brings limited improvements, yet doesn't defect the performance, compared to techniques like instance normalization.

Method	M→MS		VID→VR	
	R-1↑	mAP↑	R-1↑	mAP↑
OSNet	21.2	7.2	69.2	32.0
MSINet	**22.4**	**8.3**	**72.1**	**33.8**

Table 10. Additional Cross-domain Experiments

Method	M		VR→VID	
	R-1↑	mAP↑	R-1↑	R-5↑
SAM	**95.5**	**89.9**	**49.0**	**66.8**
SAM-softmax	95.0	89.1	48.2	65.9

Table 11. Ablation study on softmax operation.

7. More Ablation Study

Additional cross-domain experiments. We add the M→MS and VID→VR experiment results to Tab. 10. MSINet surpasses OSNet on all metrics.

Ablation study on softmax operation. SAM aligns the activation values in the feature map, where the discriminative positions are actually matched between different samples. As the "Mutual Conv" operation is non-parametric, it is not proper to apply the softmax-squeezed position attention values for direct alignment, which may result in scale inconsistency. The experiment results in Tab. 11 also suggest slight influence on this detail.

8. Limitations and Future Work

The designed interaction operations only include forward and exchange in the direct and attention forms, which restricts the size of the search space. In the future works, there are still exploration room for more elaborate and complicated interaction operations and search spaces. There are

still exploration room for more complicated search schemes and spaces.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[2] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. In *ICLR*, 2018. 2

[3] Xinchen Liu, Wu Liu, Huadong Ma, and Huiyuan Fu. Large-scale vehicle re-identification in urban surveillance videos. In *ICME*, pages 1–6, 2016. 1

[4] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *ECCV*, pages 869–884, 2016. 1

[5] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *T-MM*, 22(10):2597–2609, 2019. 1, 2

[6] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018. 1

[7] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 2

[8] Meng Zheng, Srikrishna Karanam, Ziyan Wu, and Richard J Radke. Re-identification with consistent attentive siamese networks. In *CVPR*, pages 5735–5744, 2019. 3

[9] Sanping Zhou, Fei Wang, Zeyi Huang, and Jinjun Wang. Discriminative feature learning with consistent attention regularization for person re-identification. In *ICCV*, pages 8040–8049, 2019. 3