Supplementary Material for "RGB-Infrared Cross-Modality Person Re-Identification"

Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong and Jianhuang Lai

Abstract

This supplementary material accompanies the paper "RGB-Infrared Cross-Modality Person Re-Identification". It includes more details of Section 4, as well as extra evaluations of our proposed deep zero-padding method.

1. Details of Counting Domain-Specific Nodes

In the third paragraph of Section 4.2 in the main manuscript, we quantify the number of domain-specific nodes in the trained network in our experiments.

As defined in Equation (3) in Section 3 in the main manuscript, the categorization of node types is rather strict. In the l-th layer, let $\eta_i^{(l)}$ denote the i-th node and $f_{out}(\mathbf{x}^{(0)}, i, l)$ denote the output of $\eta_i^{(l)}$ given the network input $\mathbf{x}^{(0)}$. Let $\mathbf{x}_{d1}^{(0)}$ and $\mathbf{x}_{d2}^{(0)}$ be inputs of the whole network of domain1 and domain2, respectively. The type of node $\eta_i^{(l)}$ is defined by

$$type(\eta_i^{(l)}) = \begin{cases} domain1-specific, & f_{out}(\mathbf{x}_{d2}^{(0)}, i, l) \equiv 0 \\ domain2-specific, & f_{out}(\mathbf{x}_{d1}^{(0)}, i, l) \equiv 0 \\ shared, & otherwise. \end{cases} \quad (1)$$

Since the identity sign is used here, the categorization condition is too strict in applications. So we relax the categorization condition for counting towards domain-specific nodes in application by setting a threshold T. The relaxed definition of node type is formulated as follows: for all $\mathbf{x}_{d1}^{(0)}$ and $\mathbf{x}_{d2}^{(0)}$ in our experiments,

$$type(\eta_i^{(l)}) = \begin{cases} domain1-specific, & f_{out}(\mathbf{x}_{d2}^{(0)}, i, l) < T \; and \\ & \quad f_{out}(\mathbf{x}_{d1}^{(0)}, i, l) > T \\ domain2-specific, & f_{out}(\mathbf{x}_{d1}^{(0)}, i, l) < T \; and \\ & \quad f_{out}(\mathbf{x}_{d2}^{(0)}, i, l) > T \\ shared, & otherwise. \end{cases} \quad (2)$$

Because the scales of responses on feature maps differ from layer to layer, we set $T = \alpha \; std(x_i^{(l)})$, where α is a proportion coefficient, $x_i^{(l)}$ is the output value of the i-th node in the l-th layer and $std(\cdot)$ is the standard deviation function. For an image channel in our experiments, we compute the average of all values in the feature map as the output of the node. We set $\alpha = 0.01$ and $\alpha = 0.05$ for strict and loose categorizations, respectively. The relation between the proportion of domain-specific nodes and layer depth is shown in Figure S1. Both total proportions and respective proportions of two domains are shown. With strict threshold, domain-specific nodes mainly exist in the first three layers. With loose threshold, domain-specific nodes mainly exist in the first five layers. In both cases, the network can learn more domain-specific nodes using deep zero-padding. When the threshold is loosened, the proportion of domain-specific nodes increases when using deep zero-padding, but keeps nearly unchanged when using the inputs without zero-padding.

2. Evaluation on Using Different Networks

Our deep model is based on ResNet [1] as illustrated in Section 5 in the main manuscript. Deep zero-padding has shown effectiveness on ResNet-6 in our experiments. To verify whether deep zero-padding can also work with other one-stream networks, we also evaluated our method on popular architectures AlexNet [2] and VGG-16 [3]. The results are reported in Table S1.

Generally, using deep zero-padding can improve the performance in most cases for all evaluated network architectures. The improvement is especially evident for ResNet-6.

References

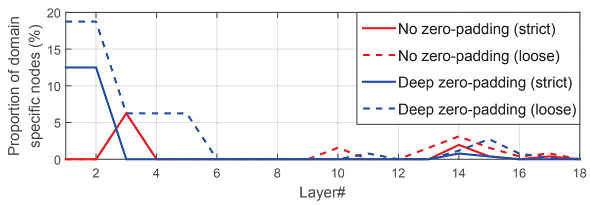
[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

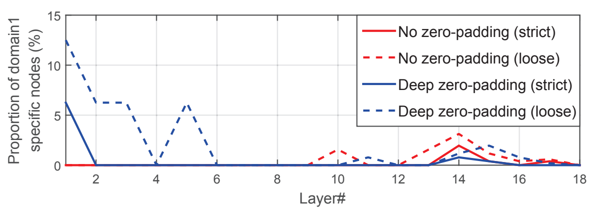
[3] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1

Table S1. Performance under all-search and indoor-search using different networks, where r1, r10, r20 denote rank-1, 10, 20 accuracies (%), respectively and mAP denotes mean average precision (%).

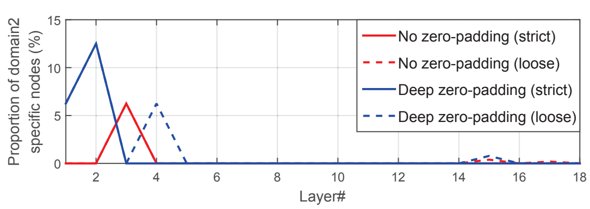
Method	Metric	All-search								Indoor-search							
		Single-shot				Multi-shot				Single-shot				Multi-shot			
		r1	r10	r20	mAP	r1	r10	r20	mAP	r1	r10	r20	mAP	r1	r10	r20	mAP
ResNet-6 (deep zero-padding)	Euclidean	**14.80**	**54.12**	**71.33**	**15.95**	**19.13**	**61.40**	**78.41**	**10.89**	**20.58**	**68.38**	**85.79**	**26.92**	**24.43**	**75.86**	**91.32**	**18.64**
ResNet-6	Euclidean	12.04	49.68	66.74	13.67	16.26	58.14	75.05	8.59	16.94	63.55	82.10	22.95	22.62	71.74	87.82	15.04
VGG-16 (deep zero-padding)	Euclidean	**9.23**	**39.14**	**55.38**	**9.60**	**11.45**	**45.50**	**62.41**	**5.93**	**11.45**	**53.18**	**73.73**	**17.20**	**14.82**	**62.01**	**80.88**	**10.13**
VGG-16	Euclidean	7.46	36.52	51.71	8.69	9.42	43.49	60.30	5.20	10.61	50.02	70.29	16.25	14.27	60.97	79.87	9.37
AlexNet (deep zero-padding)	Euclidean	**9.70**	**43.14**	**59.25**	**11.00**	**11.52**	**50.04**	**67.50**	**6.68**	**12.96**	**55.88**	75.45	**19.12**	15.41	**62.51**	**81.22**	**11.71**
AlexNet	Euclidean	9.48	41.63	57.96	10.32	11.07	49.38	66.53	6.21	12.69	55.40	**75.50**	18.67	**16.16**	61.31	79.73	11.42



(a) Proportions of domain-specific nodes



(b) Proportions of domain1-specific (RGB) nodes



(c) Proportions of domain2-specific (IR) nodes

Figure S1. Relation between proportion of domain-specific nodes and layer depth. The x-axis denotes layer depth from bottom to top of the network, and the y-axis denotes the proportion of domain-specific nodes. Generally, the proportion of domain-specific nodes using deep zero-padding is higher than that without zero-padding.