Supplementary Material for Multi-scale Matching Networks for Semantic Correspondence

Dongyang Zhao[1,2], Ziyang Song[1], Zhenghao Ji[1], Gangming Zhao[3], Weifeng Ge *[1,2], and Yizhou Yu[3]

[1]Nebula AI Group, School of Computer Science, Fudan University
[2]Shanghai Key Lab of Intelligent Information Processing
[3]Department of Computer Science, The University of Hong Kong

A. Parameter Table

Images are resized to 224×320 for all datasets both in training and testing.

		ResNet-50	ResNet-101	ResNeXt-101	ResNet-101-FCN
Conv Intra-fused	conv: output channel = 21, kernel size = 1×1 , stride = 1				
Conv Cross-fused	conv: output channel = 21, kernel size = 3×3, stride = 1, padding = 1				
Deconv	transpose conv: output channel = 21, kernel size = 4×4, stride = 2, padding = 1 , bias = False				
LSA	conv_key	conv: output channel = 10, kernel size = 1×1			
	conv_value	conv: output channel = 10, kernel size = 1×1			
	conv_query	conv: output channel = 10, kernel size = 1×1			
	conv_aggregate	conv: output channel = 21, kernel size = 1×1			
# Params	backbone	25.6×10^6	$44.5.0 \times 10^6$	88.8×10^6	54.4×10^6
	MMNet w/o backbone	4.8×10^6	10.3×10^6	10.3×10^6	10.3×10^6
	MMNet	30.4×10^6	54.8×10^6	99.1×10^6	64.7×10^6
FLOPs	backbone	11.8×10^9	22.4×10^9	47.0×10^9	127.6×10^9
	MMNet w/o backbone	3.1×10^9	4.6×10^9	4.6×10^9	12.7×10^9
	MMNet	14.9×10^9	27.0×10^9	51.6×10^9	140.3×10^9

Table 1. Implementation details of MMNet, and numbers of parameters and FLOPs introduced by different modules. In this table, *'Conv Intra-fused'* and *'Conv Cross-fused'* denote the convolution operation in the intra-scale and cross-scale feature enhancements separately, *'Deconv'* indicates the deconvolution operation to upscale the feature maps during the cross-scale feature enhancement, and *'LSA'* is short for the local self attention module used at the end of the intra-scale feature enhancement.

Analysis. MMNet introduces additional parameters only in its feature enhancement module. We follow BDCN [1] to set the parameters of scale enhancement modules. Each scale enhancement module contains four 3×3 convolution operations with dilation 1, 4, 8, 12 respectively. The output channel numbers of these convolution operations are set to 32. The output channel numbers of other convolution and deconvolution operations are set to 21 to reduce the computational cost. We compare the additional parameters and FLOPs introduced by MMNet with four different backbones including ResNet-50, ResNet-101, ResNeXt-101 and ResNet-101-FCN. The additional parameters are no larger than 25% of that of any backbone. For the FLOPs, we add at most 26.3% computational cost when using ResNet-50 as the backbone.

B. Additional Results

Results with other backbones. We adapt our MMNet design with other backbones: VGG-16 and DeepLab-V3, the results are listed in 2. As can be seen, our model with ResNet-101-FCN backbone performs significantly best than others.

*Corresponding author: wfge@fudan.edu.cn

Methods	PF-PASCAL		
	0.05	0.1	0.15
$\text{MMNet}_{\text{ResNet-101-FCN}}$	81.1	91.6	95.9
$\text{MMNet}_{\text{VGG-16}}$	69.5	82.0	88.4
$\text{MMNet}_{\text{DeepLabV3-ResNet101}}$	73.3	85.3	92.3

Table 2. Experiments on different backbones. All experiments is conducted on PF-PASCAL.

Figure 1. Key-point matching results on SPair-71k dataset [3] compared with SCOT [2] and DHPF [4]. The odd rows are the source images, and the even rows are the target images. Destination key points are denoted with crosses.

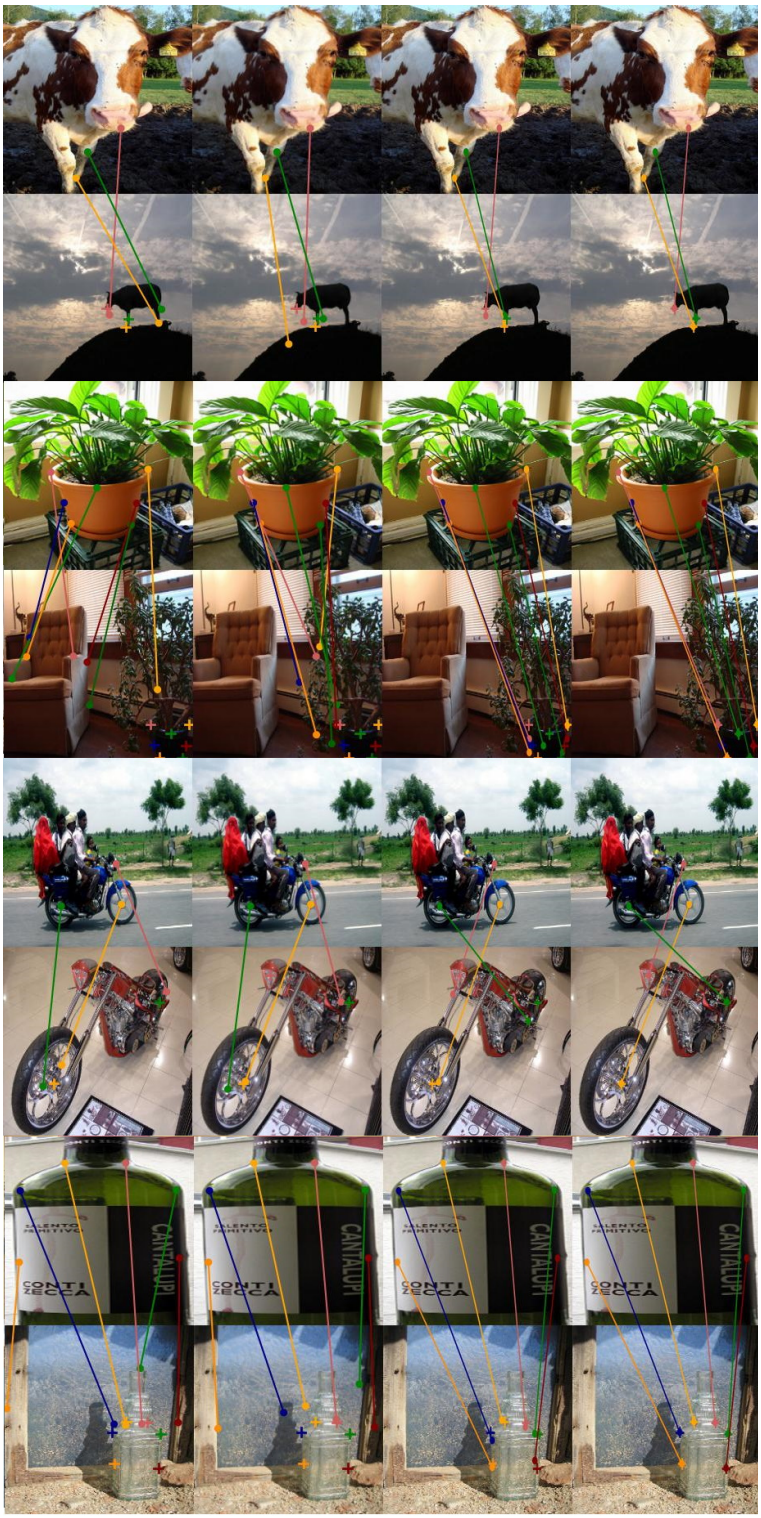

Figure 2. Key-point matching results on SPair-71k dataset [3] compared with SCOT [2] and DHPF [4]. The odd rows are the source images, and the even rows are the target images. Destination key points are denoted with crosses.

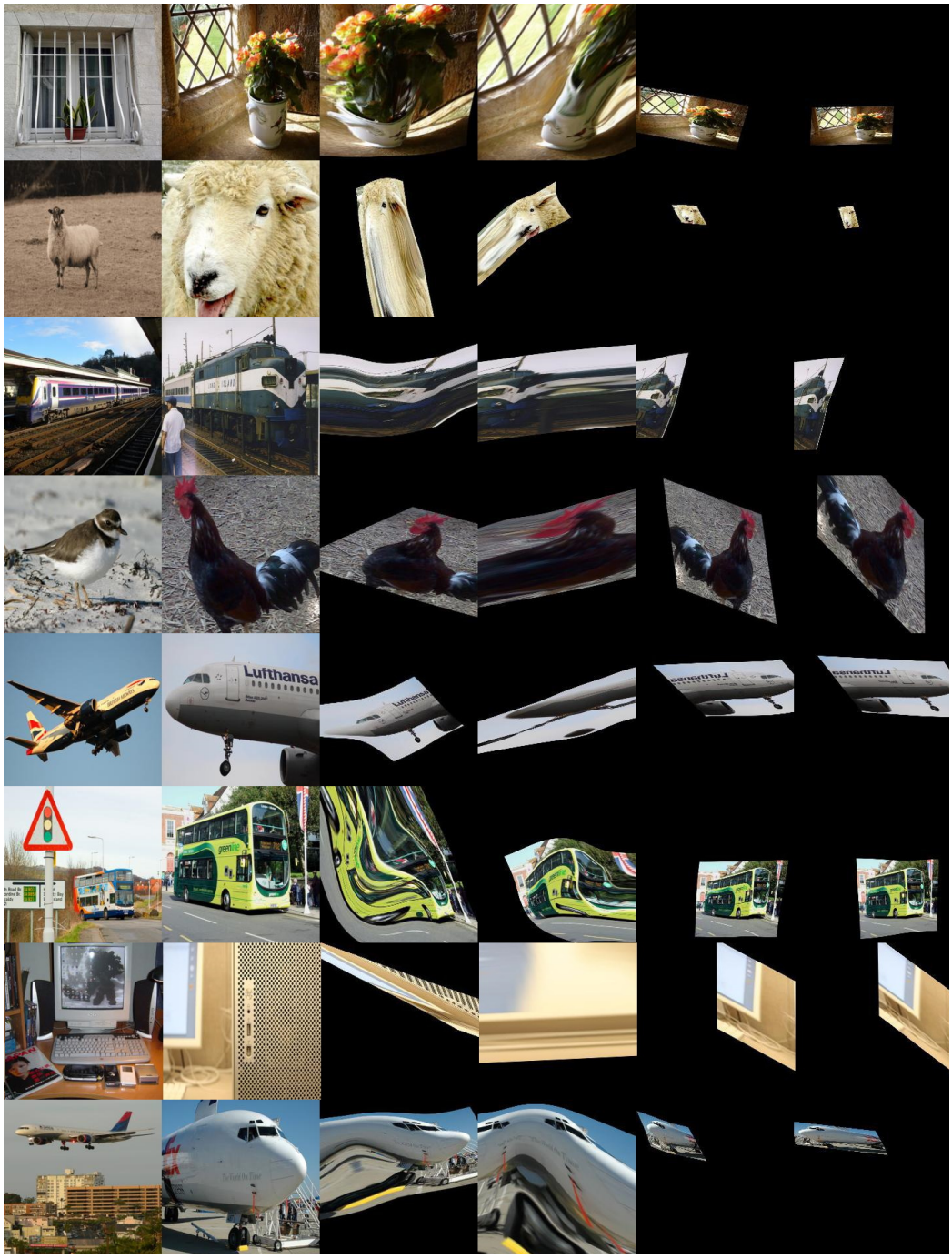

Figure 3. Warped images by thin-plate splines with the predicted key point pairs on SPair-71k dataset [3] compared with SCOT [2] and DHPF [4]

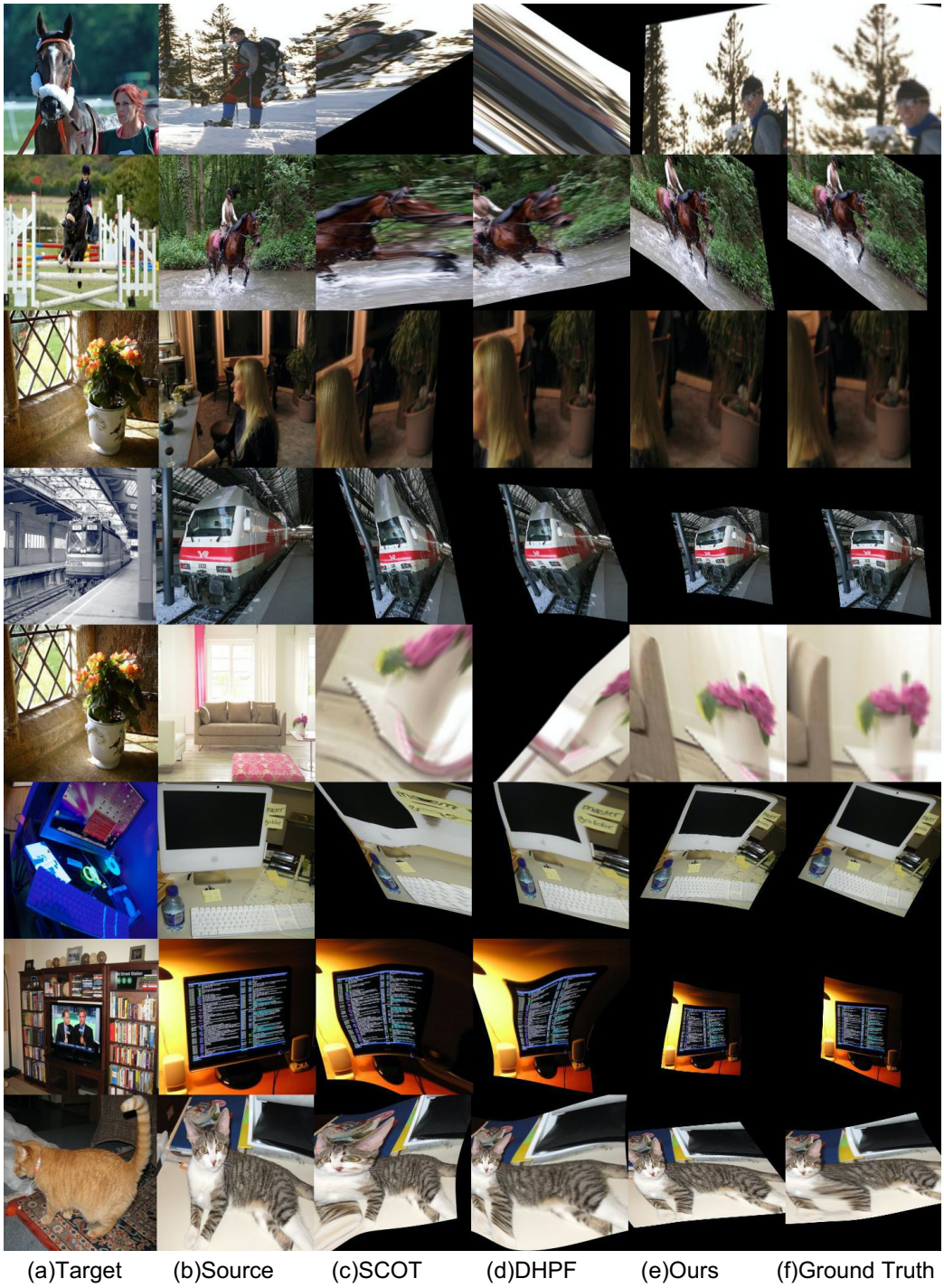

Figure 4. Warped images by thin-plate splines with the predicted key point pairs on SPair-71k dataset [3] compared with SCOT [2] and DHPF [4]

References

[1] Jianzhong He, Shiliang Zhang, Ming Yang, Yanhu Shan, and Tiejun Huang. Bi-directional cascade network for perceptual edge detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3828–3837, 2019. 1

[2] Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic correspondence as an optimal transport problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4463–4472, 2020. 3, 4, 5, 6

[3] Juhong Min, Jongmin Lee, J. Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *ArXiv*, abs/1908.10543, 2019. 3, 4, 5, 6

[4] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Learning to compose hypercolumns for visual correspondence. In *ECCV 2020-16th European Conference on Computer Vision*. Springer, 2020. 3, 4, 5, 6