Deep Adversarial Decomposition: A Unified Framework for Separating Superimposed Images

Supplementary Material

Zhengxia Zou[1], Sen Lei[2], Tianyang Shi[3], Zhenwei Shi[2], Jieping Ye[1,4]
[1]University of Michigan, Ann Arbor, [2]Beihang University,
[3]NetEase Fuxi AI Lab, [4]Didi Chuxing

1 Training details

In our default settings, we train our model for 200 epochs by using the Adam optimizer [6] with batch_size $= 2$. For the first 100 epochs, we set learning_rate $= 0.0001$. For the rest epochs, we reduce the learning rate to its 1/10. We set $\beta_C = \beta_M = 0$ for the first 10 epochs to stabilize the training and then set $\beta_C = \beta_M = 0.001$ for the rest epochs. We do not use batch-normalization or dropout layers in G as we found it may introduce unexpected artifacts.

As is suggested by I. Goodfellow *et al.*[2], instead of training G to minimize $\log(1 - D(G(\cdot)))$, in practice, we try to maximize $\log D(G(\cdot))$. This is because in early stage of learning, $\log(1 - D(G(\cdot)))$ tends to saturate. This revision on objective provides much stronger gradients early in learning.

2 Configurations of our Networks

We build our separator G by following the configuration of the UNet [10]. For input images of three different sizes, i.e., 128x128, 256x256, and 512x512, we set the layer number of our separator to 14, 16, 18, respectively. We add skip connections to our separator between the layer i and layer $n - i$ for learning both high-level semantics and low-level details. We remove the nonlinear activation on the last layer of our separator since we found it may slow-down the convergence.

We build our discriminators by following the configurations of Pix2Pix [4]. We build our D_C, D_{M1} and D_{M2} as three standard FCNs with 4, 3, and 3 convolutional layers. The perceptive fields of D_{M1} and D_{M1} are set to $N = 30$. We resize the input of D_C to a relatively small size, e.g., 64×64, and set the receptive field size larger than this size to capture the semantics of the whole image instead of adding more layers or using larger pooling/convolutional strides.

Suppose "CDk" represents a down-sampling convolution layer with k filters, spatial_size=4×4 and stride=2; "CUk" represents a up-sampling fractional-strided convolution layer (a.k.a. the transposed convolution) [15] with k filters, spatial_size=4×4 and stride=1/2; "BN" represents a batch-normalization layer; "xk" means we repeat the module x for k times. All ReLUs in the down sampling layers of any of our networks are set to leaky ReLUs with slope=0.2, while those in the up-sampling layers are set to standard ones.

The architectures of our separators are as follows:

UNet128: CD64-CD128-CD256-CD512x4-CU512-CU1024x3-CU512-CU256-CU128.
UNet256: CD64-CD128-CD256-CD512x5-CU512-CU1024x4-CU512-CU256-CU128.
UNet512: CD64-CD128-CD256-CD512x6-CU512-CU1024x5-CU512-CU256-CU128.

The architectures of our discriminators are as follows:

Critic D_C: CD64-CD128-BN-CD256-BN-CD512-BN-CD512.
Distriminator D_{Mi} (i=1,2): CD64-CD128-BN-CD256-BN-CD256.

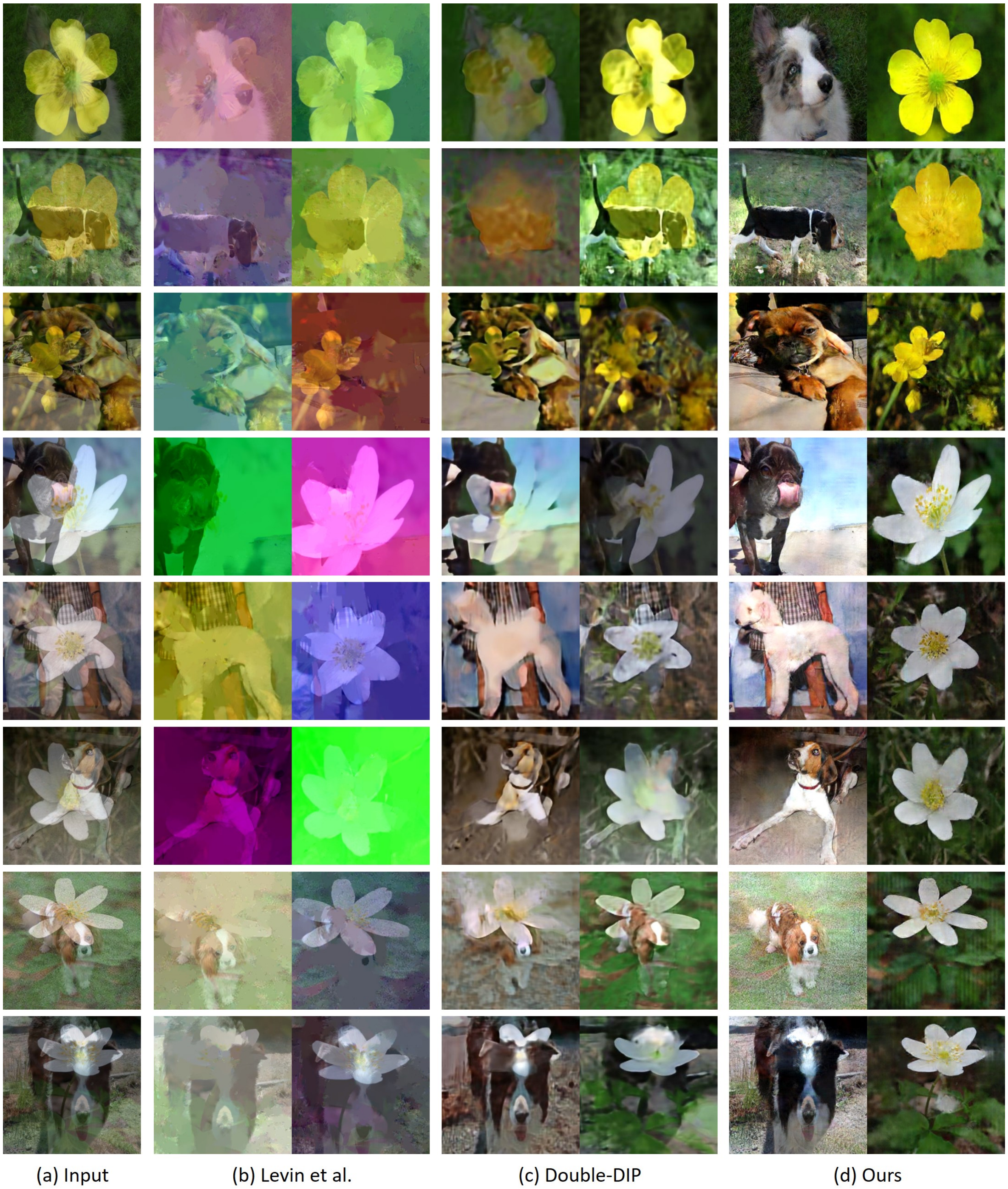

Figure 1: More comparison results on single mixed image separation: (a) input mixed image, (b) the method of Levin *et al.*[7], (c) Double-DIP [1], and (d) our method. Datasets: Stanford-Dogs [5] + VGG-Flowers [8].

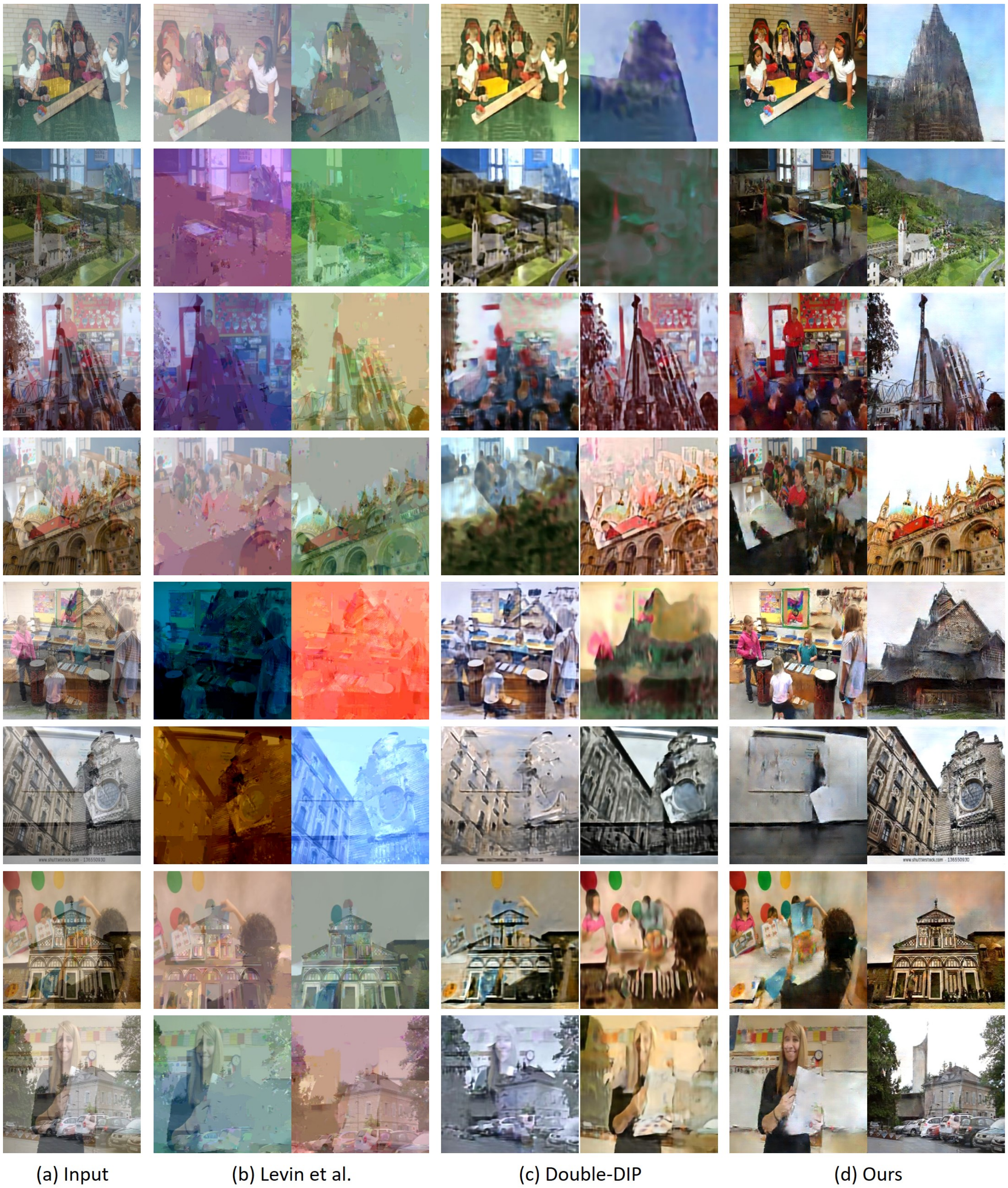

Figure 2: More comparison results on single mixed image separation: (a) input mixed image, (b) the method of Levin *et al*.[7], (c) Double-DIP [1], and (d) our method. Datasets: LSUN Classroom + LSUN Church [14].

Figure 3: More comparison results of different reflection removal methods: BDN [13] (ECCV'18), RmNet [12] (CVPR'19), and our method, on some real-world reflection images from the dataset [16].

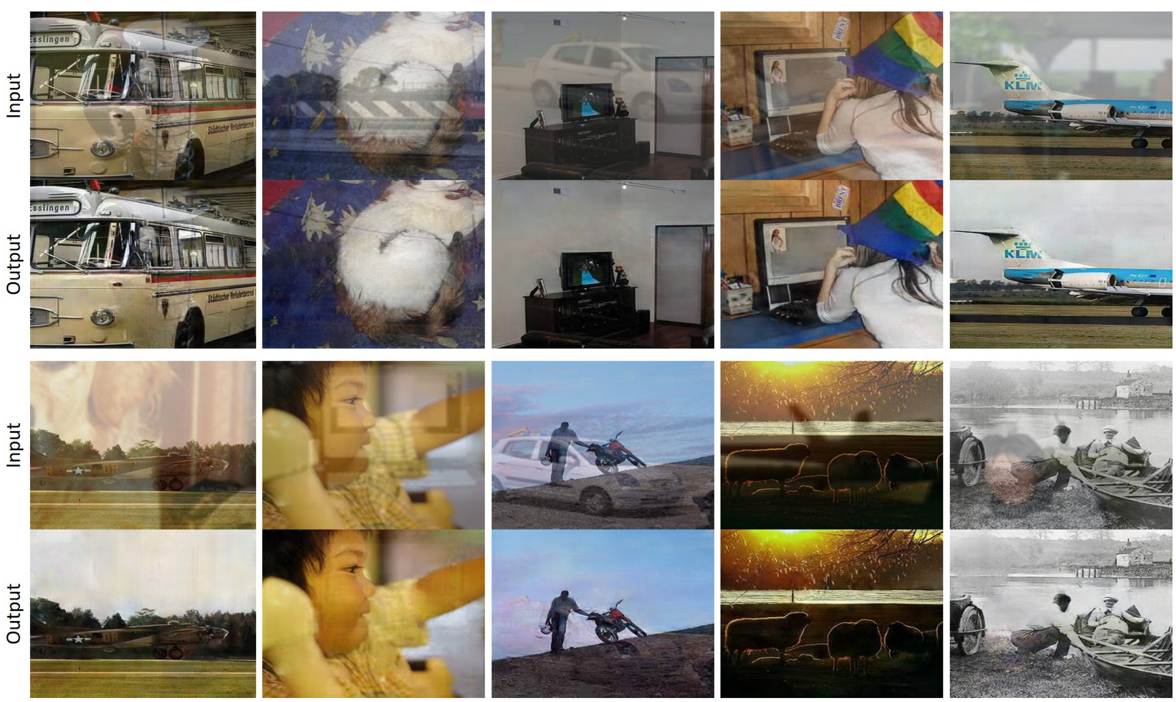

Figure 4: More examples of the reflection removal results with our method on the BDN dataset [13].

Figure 5: More comparison results between our method and DSC (TPAMI19) [3] on the shadow removal dataset ISTD [11].

Figure 6: More comparison results between our method and DSC (TPAMI19) [3] on the shadow removal dataset SRD [9].

References

[1] Yosef Gandelsman, Assaf Shocher, and Michal Irani. ”double-dip”: Unsupervised image decomposition via coupled deep-image-priors. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[3] X Hu, CW Fu, L Zhu, J Qin, and PA Heng. Direction-aware spatial context features for shadow detection and removal. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[5] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2, 2011.

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[7] Anat Levin and Yair Weiss. User assisted separation of reflections from a single image using a sparsity prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1647–1654, 2007.

[8] M-E Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1447–1454. IEEE, 2006.

[9] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson W. H. Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[11] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[12] Qiang Wen, Yinjie Tan, Jing Qin, Wenxi Liu, Guoqiang Han, and Shengfeng He. Single image reflection removal beyond linearity. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[13] Jie Yang, Dong Gong, Lingqiao Liu, and Qinfeng Shi. Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[14] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[15] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[16] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.