CARAFE: Content-Aware ReAssembly of FEatures Supplementary Materials

Jiaqi Wang[1] Kai Chen[1] Rui Xu[1] Ziwei Liu[1] Chen Change Loy[2] Dahua Lin[1]
[1]CUHK - SenseTime Joint Lab, The Chinese University of Hong Kong
[2]Nanyang Technological University
{wj017,ck015,xr018,dhlin}@ie.cuhk.edu.hk zwliu.hust@gmail.com ccloy@ntu.edu.sg

1. Experimental Settings

Object Detection and Instance Segmentation. We evaluate CARAFE on Faster RCNN [11] and Mask RCNN [4] with the ResNet-50 backbone [5]. FPN [8] is used for these methods. In both training and inference, we resize an input image such that its shorter edge has 800 pixels or longer edge has 1333 pixels without changing its aspect ratio. We adopt synchronized SGD with an initial learning rate of 0.02, a momentum of 0.9 and a weight decay of 0.0001. We use a batchsize of 16 over 8 GPUs (2 images per GPU). Following the 1x training schedule as Detectron [3] and MMDetection [1], we train 12 epochs in total and decrease the learning rate by a factor of 0.1 at epoch 8 and 11.

Semantic Segmentation. We use the official implementation of UperNet[1] [12] with the ResNet-50 backbone. During the training, an input image is resized such that the size of its shorter edge is randomly selected from {300, 375, 450, 525, 600}. In inference, we apply the single scale testing for a fair comparison and the shorter edge of an image is set to 450 pixels. The maximum length of the longer edge of an image is set to 1200 in both training and inference. We adopt synchronized SGD with an initial learning rate of 0.02, a momentum of 0.9 and a weight decay of 0.0001. We use a batchsize of 16 over 8 GPUs (2 images per GPU), and synchronized batch normalization is adopted as a common practice in semantic segmentation. Following [2], the 'poly' learning rate policy in which the learning rate of current iteration equals to the initial learning rate multiplying $(1 - iter/max\_iter)^{power}$ is adopted. We set $power$ to 0.9 and train 20 epochs in total.

Image Inpainting. We employ the generator and discriminator networks from Global&Local [6] as the baseline. Our generator takes a 256×256 image $\mathbf{x}$ with masked region M as input and produces a 256×256 prediction of the missing region $\hat{\mathbf{y}}$ as output. Then we combine the predicted image with the input by $\mathbf{y} = (1 - M) \odot \mathbf{x} + M \odot \hat{\mathbf{y}}$. Finally, the combined output $\mathbf{y}$ is fed into the discriminator. We apply a simple modification to the baseline model to achieve better generation quality. Compared to the original model that employs two discriminators, we employ only one PatchGAN-style discriminator[7] on the inpainted region. This modification can achieve better image quality.

For a fair comparison and taking real-world application into consideration, we use the free-form masks introduced by [13] as the binary mask M. For Partial Conv [10], we just substitute the convolution layers with the official Partial Conv module in our generator. During training, Adam solver with learning rate 0.0001 is adopted where $\beta_1 = 0.5$ and $\beta_2 = 0.9$. Training batch size is 32. The input and output are linearly scaled within range $[-1, 1]$.

[1] https://github.com/CSAILVision/semantic-segmentation-pytorch

2. Visualization of CARAFE

We demonstrate how CARAFE performs content-aware reassembly with more examples in Figure 1. Red units are reassembled into the green center unit by CARAFE in the top-down pathway of a FPN structure.

3. Visual Results Comparison

Object Detection and Instance Segmentation. As illustrated in Figure 2, we provide more object detection and instance segmentation results comparison between Mask RCNN baseline and Mask RCNN w/ CARAFE on COCO [9] 2017 val.

Semantic Segmentation. We compare the semantic segmentation results between UperNet baseline and UperNet w/ CARAFE on ADE20k [15] val in Figure 3.

Image Inpainting. Comparison of image inpainting results between Global&Local baseline and Global&Local w/ CARAFE on Places[14] val is shown in Figure 4.

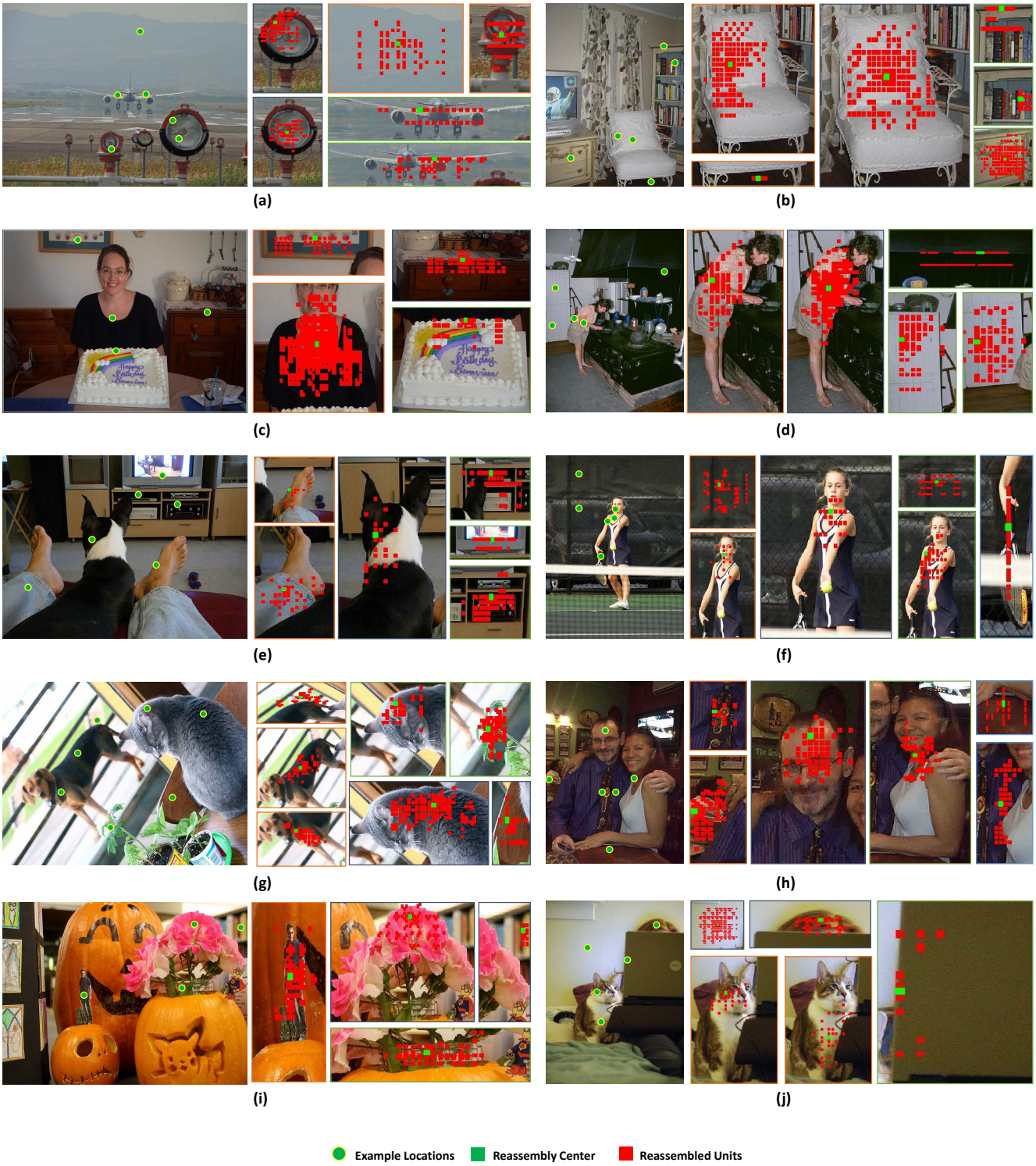

Figure 1: CARAFE performs content-aware reassembly when upsampling a feature map. Red units are reassembled into the green center unit by CARAFE in the top-down pathway of a FPN structure.

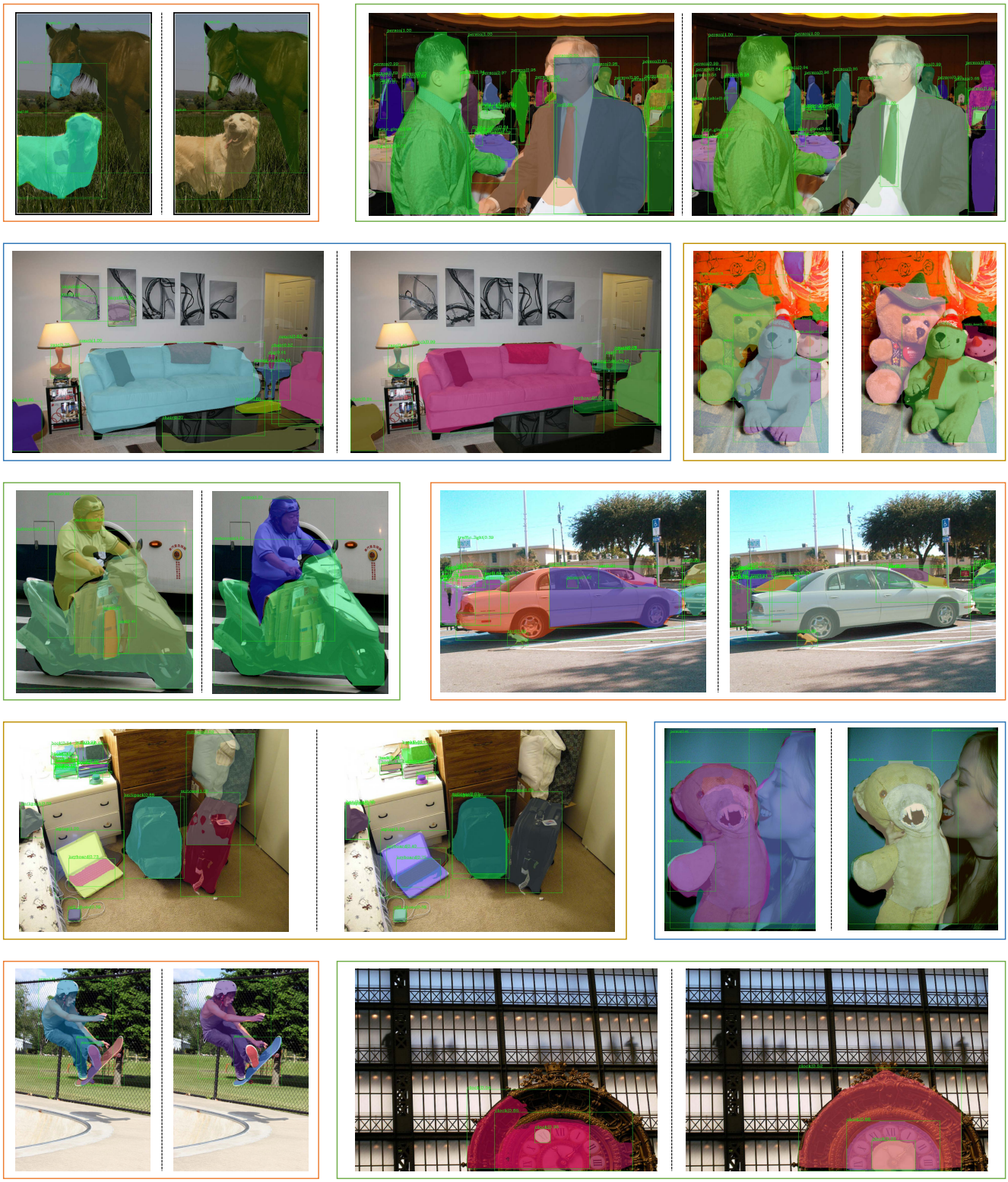

Figure 2: More comparison of object detection and instance segmentation results between Mask RCNN [4] baseline (left to the dash line) and Mask RCNN w/ CARAFE (right to the dash line) on COCO 2017 val.

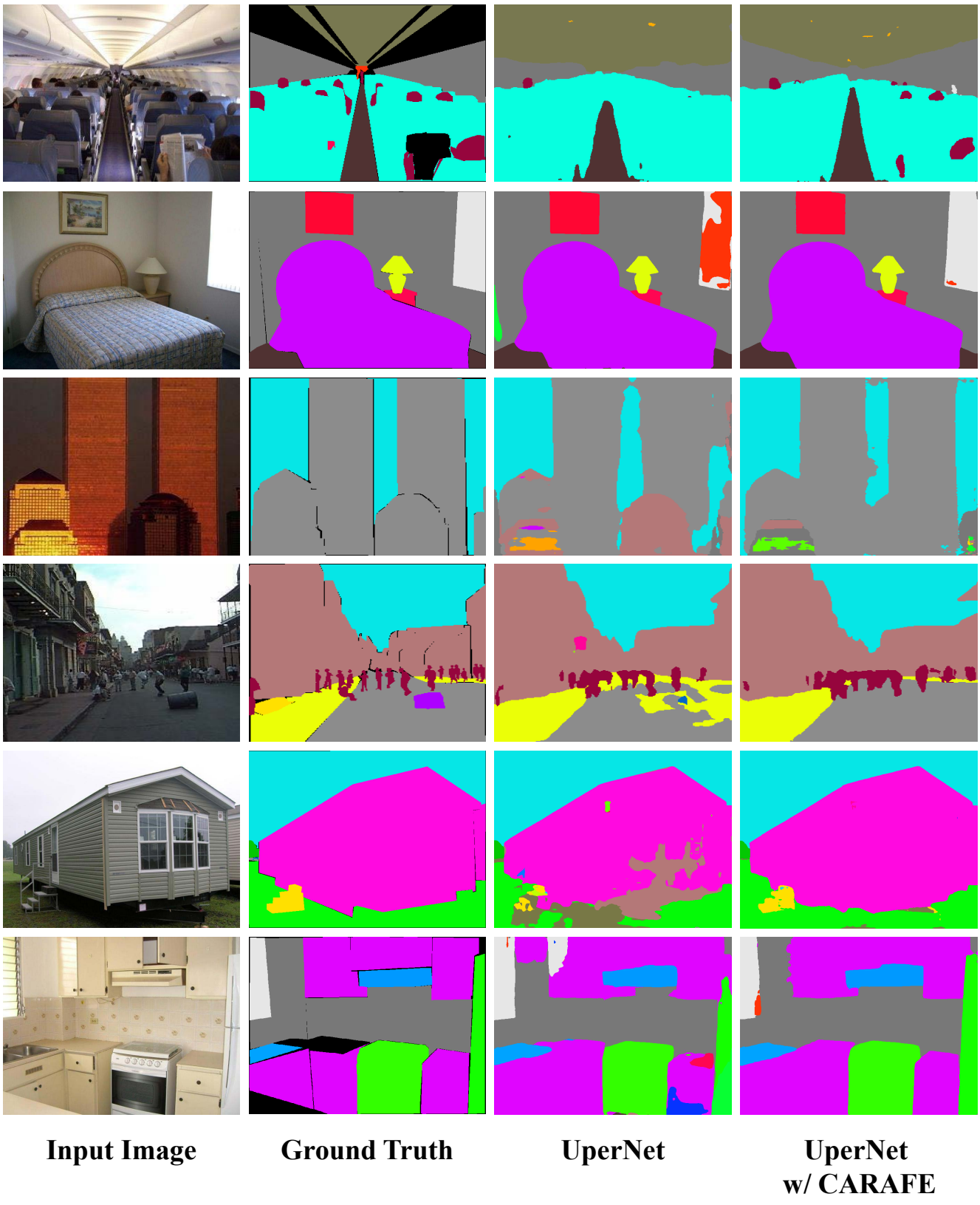

Figure 3: Comparison of semantic segmentation results between UperNet [12] baseline and UperNet w/ CARAFE on ADE20k val. Columns from left to right correspond to the input image, ground truth, baseline results and CARAFE results, respectively.

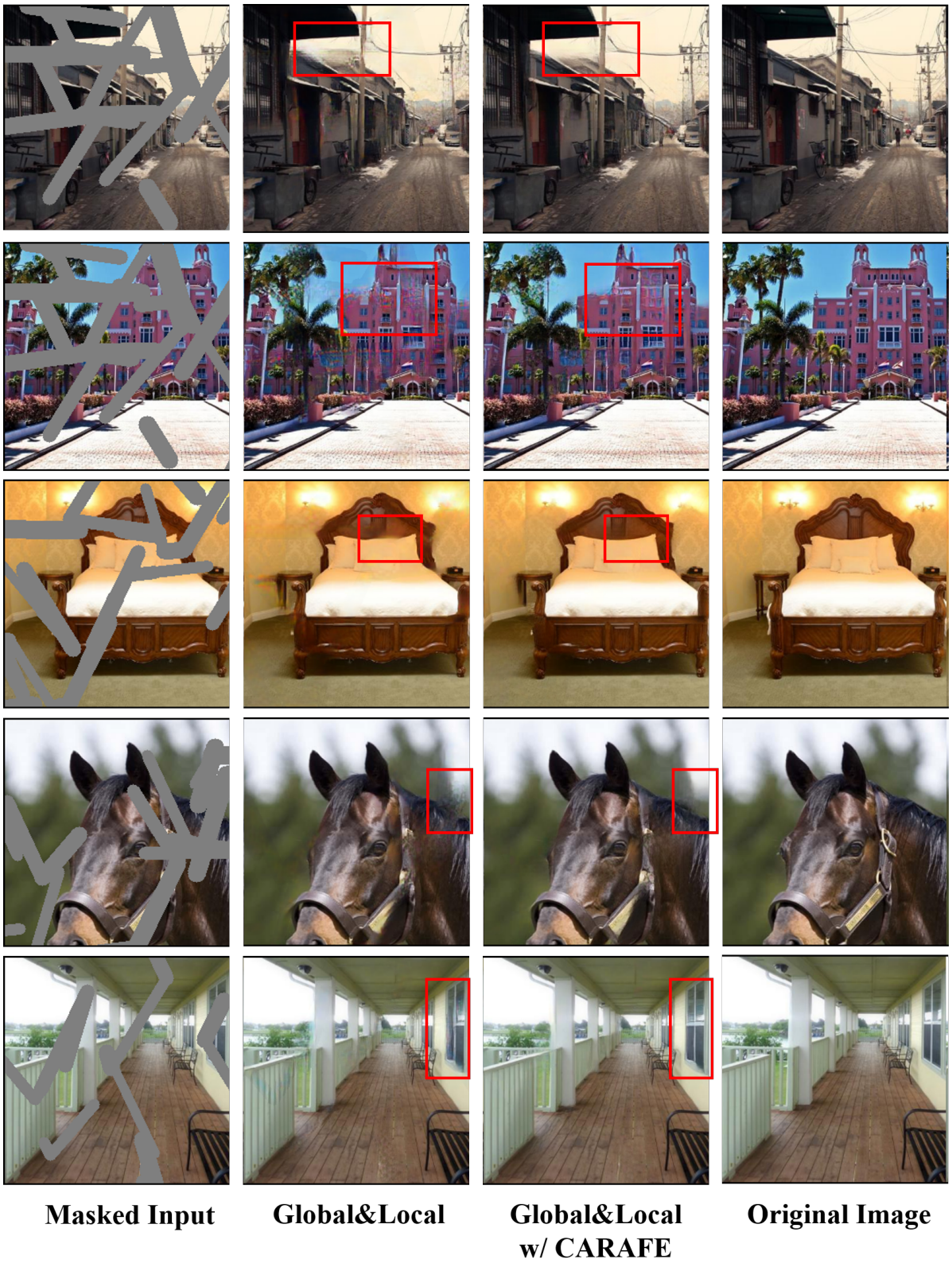

Figure 4: Comparison of image inpainting results between Global&Local [6] baseline and Global&Local w/ CARAFE on Places val. Columns from left to right correspond to the masked input, baseline results, CARAFE results and original image, respectively.

References

[1] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.

[3] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. https://github.com/facebookresearch/detectron, 2018.

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision*, 2017.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[6] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics*, 36(4):107, 2017.

[7] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, 2016.

[8] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 2014.

[10] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *European Conference on Computer Vision*, 2018.

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015.

[12] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *European Conference on Computer Vision*, 2018.

[13] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. *arXiv preprint arXiv:1806.03589*, 2018.

[14] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[15] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.