Interactive Image Manipulation with Complex Text Instructions (Supplementary Material)

Abstract

This supplementary material includes additional results and discussion of technical details that were excluded from the main submission due to the limited space. All the results are obtained with exactly the same methodology as described in the main manuscript. The process and results of the proposed image manipulation algorithm are also shown in a video. We strongly encourage the readers to check the accompanied video.

1. Additional results

Fig 1 and Fig 2 show the additional results on Caltech-UCSD Birds200-2011 (CUB) dataset [6], and Microsoft Common Objects in Context (MS COCO) datasets [3], respectively. In order to fairly compare the performance with other competing models, the automatically generated semantic map is directly used in this work. The result of this work can be further improved by refining the boundaries of object through user interface.

As shown in Fig 1, the quality of the generated images from existing works SISGAN [1], TAGAN [4] is not very high, and the boundaries of the object is not clear. Although ManiGAN [2] can generate high-quality images, the text-irrelevant contents are changed randomly and some details such as texture in the feather are lost. For our work, by applying segmentation and super resolution models, we can generate high-quality images with many details while keeping the text-irrelevant contents. Fig 2 shows the comparison results on the more complicated COCO dataset [3]. For some cases, such as the images with the input text of black cow, orange wall, and dry grass, ManiGAN [2] has the serious color overflow problem. Our work can accurately change the color of content with clear boundaries. In addition, we tackle the challenging task of not only changing the appearance information, but also resizing/removing the entity. Fig 3 shows the results of replacing the image background, changing the color, and resizing the entity. By adding the user interface, the work is able to support real-time interactive editing. As shown in Fig 4, users can keep changing the size and color of the entity.

2. Discussion of technical details

We applied two different inpainting models, Deep Image Prior (DIP) [5] and CRA [7] for two different tasks. One task is to absorb the color difference when merging the edited text-relevant contents and the original text-irrelevant contents. The other task is to tackle more complicated problems, including removing the entity, combining the resized entity, and changing the background.As shown in Fig 5, for merging an edited object, the results of applying DIP [5] are better than that of CRA [7]. For the more complicated tasks such as removing the object, CRA outperforms DIP. Based on these results, when the size of entity is not changed, DIP is applied for the merging task. Otherwise, CRA is employed for removing the entity, combining the resized entity, and changing the background.

References

[1] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE international conference on computer vision*, pages 5706–5714, 2017.

[2] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7880–7889, 2020.

[3] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[4] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: manipulating images with natural language. *Advances in neural information processing systems*, 31, 2018.

[5] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018.

[6] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

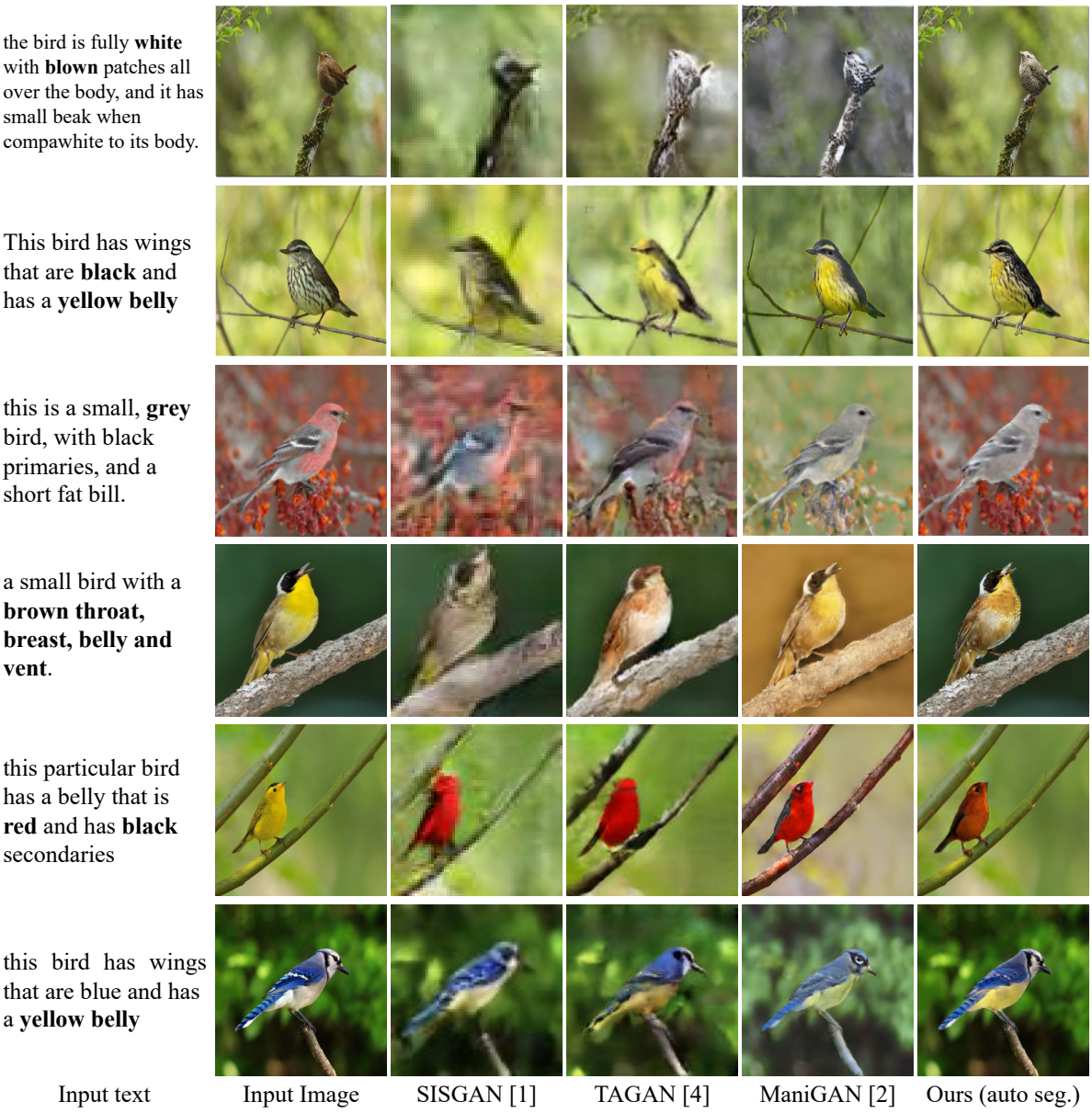

Figure 1. Additional Results on CUB dataset. To fairly compare the performance with other competing models, the automatically generated semantic map is directly used in this work. The result of this work can be further improved by refining the boundaries of object through user interface.

[7] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7508–7517, 2020.

Figure 2. Additional Results on COCO dataset. To fairly compare the performance with MainGAN, the automatically generated semantic map is directly used in this work. The result of this work can be further improved by refining the boundaries of object through user interface.

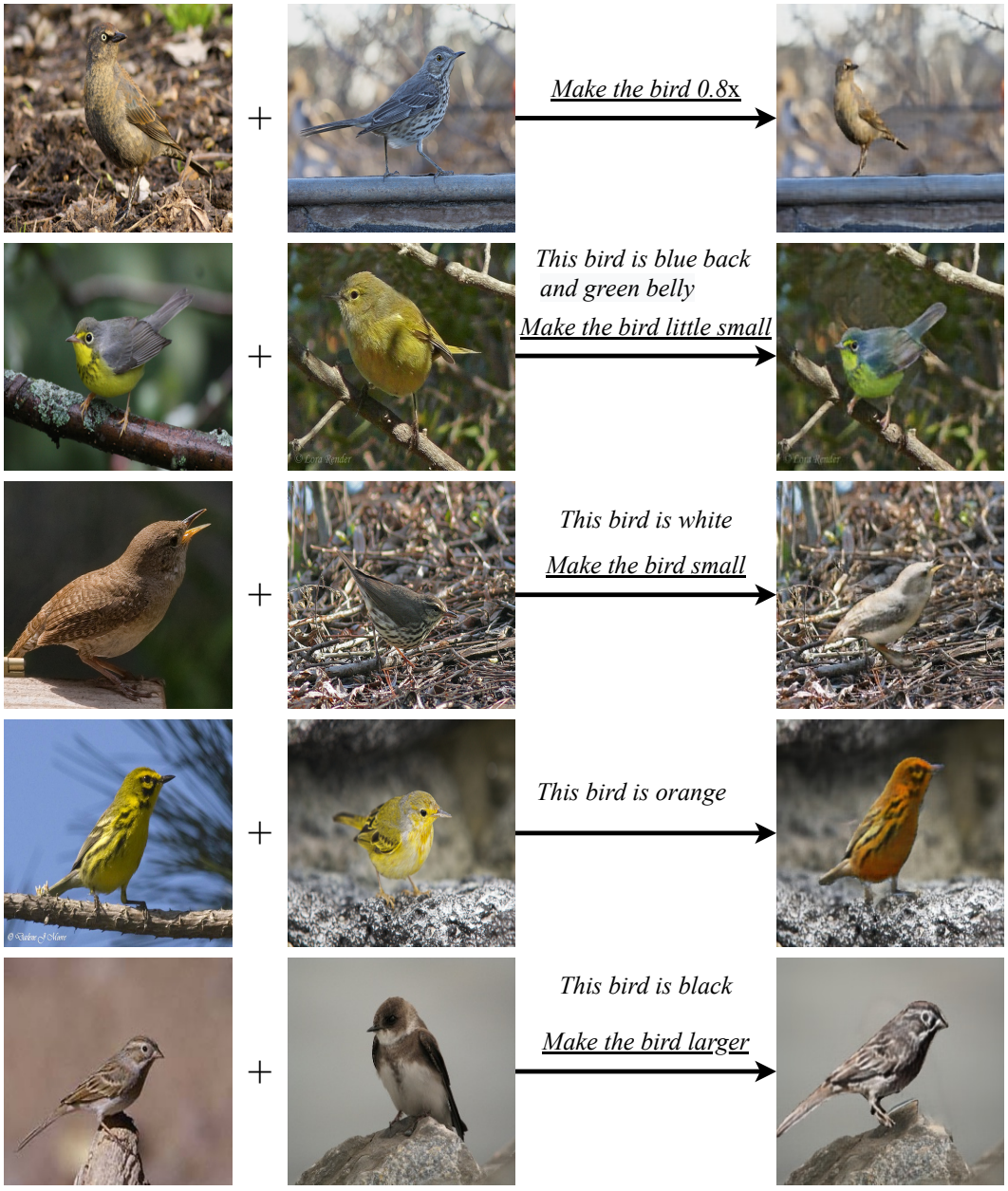

Figure 3. Result of image manipulation

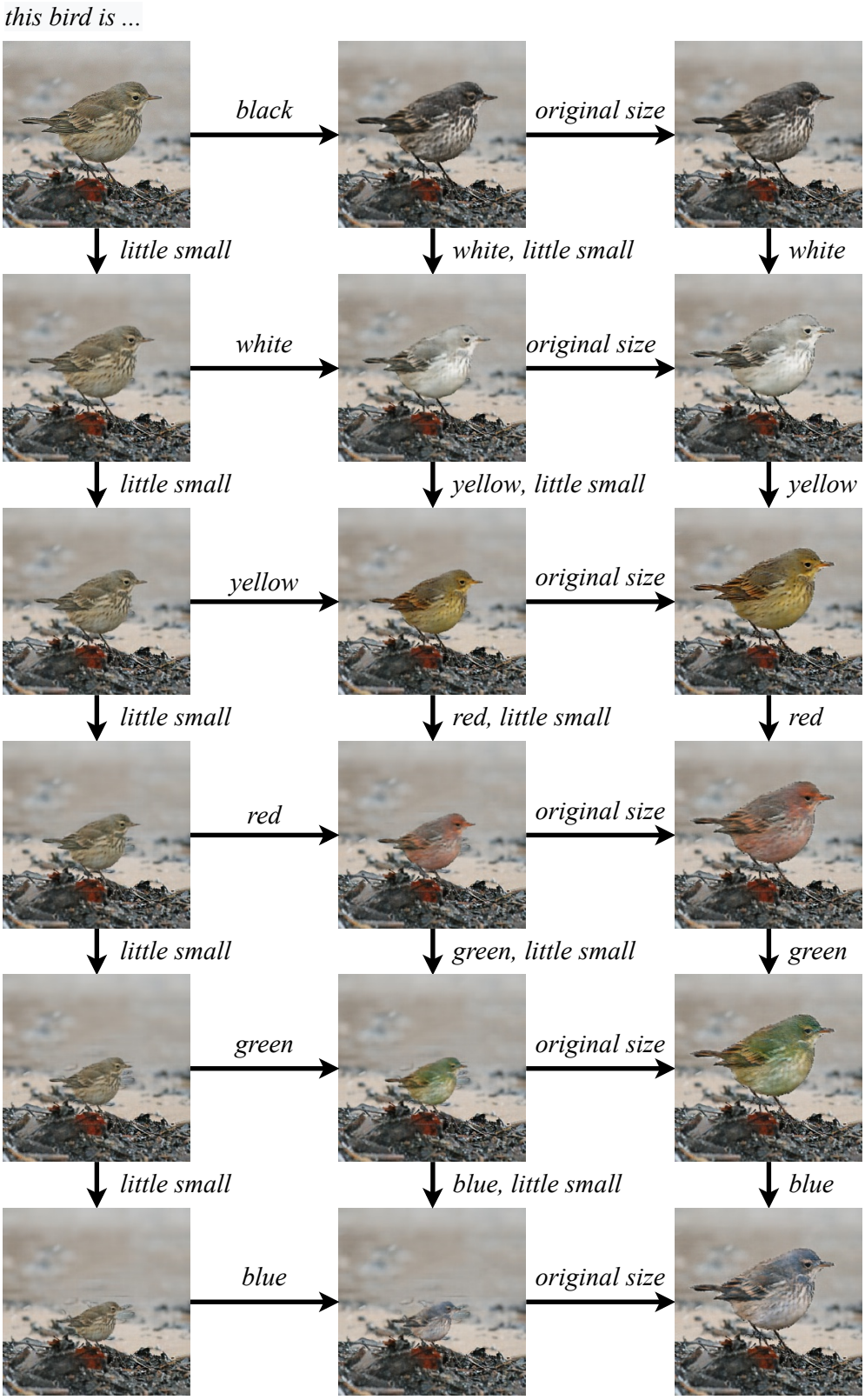

Figure 4. The result of interactive manipulation.

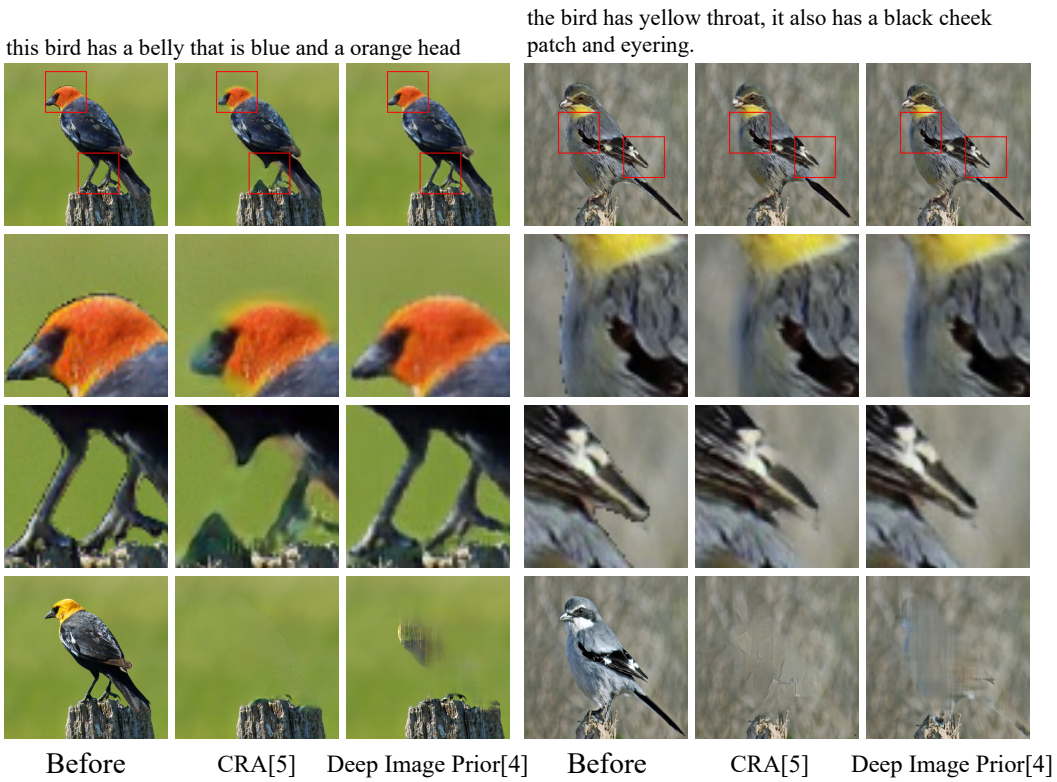

Figure 5. Comparison among before inpainting, CRA and Deep Image Prior