Supplementary Material: Efficient Layout-Guided Image Inpainting for Mobile Use

Wenbo Li, Yi Wei, Yilin Shen, Hongxia Jin
Samsung Research America
{wenbo.li1,yi.wei1,yilin.shen,hongxia.jin}@samsung.com

In this supplementary material, we use green color, *e.g.*, Fig. 3 and §3.2, to denote the figures or sections in the submitted paper. We use red color, *e.g.*, Fig. 1 and Table 1 to denote the tables or figures in the supplementary material.

1. Architecture of Refinement Model

In Table 1, we present the architecture details of the refinement model. As mentioned in §3.2 of the paper, the design of refinement model is inspired by the guided upsampling model proposed in [2]. We present the comparison between the architecture of our refinement model and that of the guided upsampling model in Fig. 1. Specifically, they have the following major differences. First, our refinement model uses the Top-X attention, and the guided upsampling model uses the vanilla contextual attention. Second, the positions where the attention mechanism is applied are different. As shown in Fig. 1 (a), in our refinement model, we place the attention mechanism between the encoder and decoder. As shown in Fig. 1 (b), in the guided upsampling model, Zeng *et al*. [2] place the attention mechanism after the decoder. Third, the guided upsampling model uses the skip connections for encoder-decoder architecture, while there are no skip connections in the refinement model.

In order to demonstrate the meaningfulness of adjusting the position of attention mechanism and removing the skip connections when designing our refinement model, we compare our method with a baseline built upon the guided upsampling model. Specifically, in order to create such a baseline, we replace the vanilla contextual attention in the guided upsampling model with the Top-X attention. The quantitative comparison and the storage and computational cost comparison are presented in Table 2 and 3, respectively. It is notable that our method achieves obviously better quantitative performance at much lower computational and storage costs.

In Fig. 2, we also show the qualitative comparison between our refinement model and the guided upsampling model. We can observe that the guided upsampling model yields obvious artifacts. We argue that this is because the

Table 1. Detailed architecture of the refinement model (§3.2 of the paper), which is corresponding to Fig. 4 in the paper. Each row describes the hyper-parameters of a gated convolutional layer. Above the "Top-X attention", we present the details of shared encoder in the refinement model. Below the "Top-X attention", we present the details of decoder in the refinement model.

Channel number	Kernel size	Stride	Activation
36	5	1	ELU
72	3	2	ELU
72	3	1	ELU
144	3	2	ELU
Top-X attention			
2× Nearest upsample			
72	1	1	ELU
72	3	1	ELU
2× Nearest upsample			
36	1	1	ELU
36	3	1	ELU
3	3	1	ELU
3	3	1	Tanh

Table 2. The quantitative experiments. ↑ (↓) means the higher (lower), the better. The best performances are highlighted in **bold**.

Method	FID ↓	SSIM ↑
GuidedUpsample	12.30	0.9574
Ours	**11.79**	**0.9587**

Table 3. Storage and computational costs. "M" and "G" represent the million and giga, respectively. The best performances are highlighted in **bold**.

Method	Params (M)	MACs (G)
GuidedUpsample	1.60	14.45
Ours	**0.48**	**8.20**

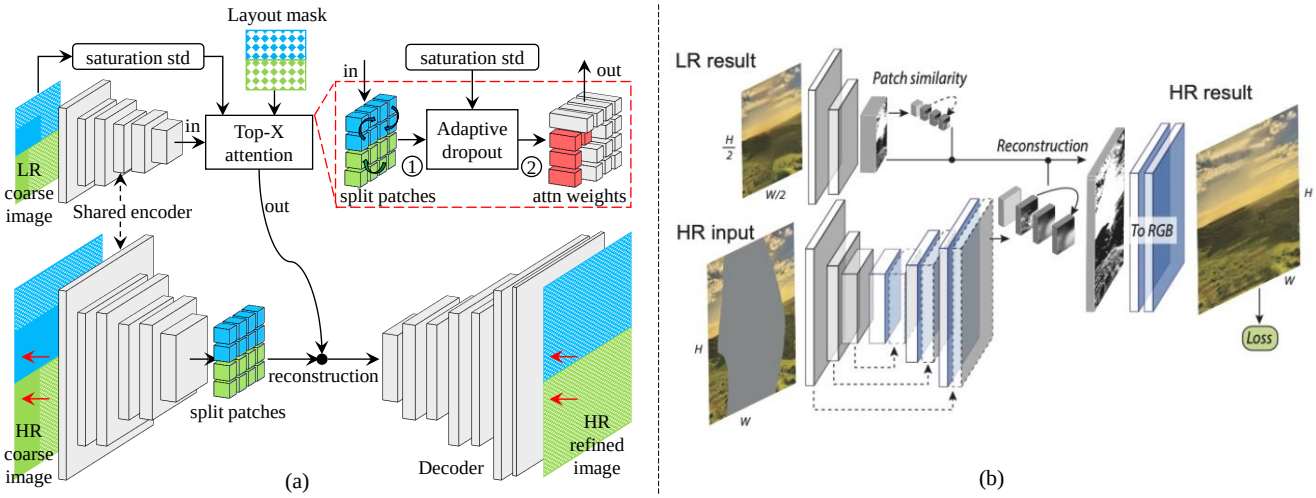

Figure 1. The architecture comparison between *(a)* our refinement model and *(b)* guided upsampling model in [2].

Table 4. The quantitative experiments. ↑ (↓) means the higher (lower), the better. The best performances are highlighted in **bold**.

Method	FID ↓	SSIM ↑
Ours w/ GAN loss	12.20	0.9585
Ours	**11.79**	**0.9587**

position where the attention mechanism is applied is too close to the output layer, which causes insufficient capacity to smooth the restored textures. It can also be observed that though our refinement model does not use the skip connections as in the guided upsampling model, we can still achieve very sharp image quality. The reduction of skip connections help reduce the storage and computational cost significantly.

2. Training

Unlike some other works, we do not use the adversarial loss for training because we observe that the adversarial loss causes the over-saturation artifacts, especially for regions with balanced RGB distribution, *e.g.*, gray-scale images. We show the quantitative comparison and the qualitative comparison in Table 4 and Fig. 3, respectively. Specifically, we directly employ the GAN loss in [1] for training.

3. Supplemental Qualitative Results

We present the output of each sub-model in our method in Fig. 4. In Fig. 5, we supplement Fig. 4 (b) of the paper with additional images which show the comparison between the contextual attention and our Top-X attention. In Fig. 6, we show the visual effects of the layout-guided diffusion model brought to our method. In Fig. 7 and Fig. 8, we supplement Fig. 5 of the paper with additional qualitative comparison among the compared methods. In Fig. 9, we show the high-resolution object removal results of our method. In Fig. 10, we show the results when applying our method to object contour adjustment. In Fig. 11, we show the results when applying our method to template based content creation.

References

[1] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Free-form image inpainting with gated convolution. In *ICCV*, 2019. 2

[2] Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. High-resolution image inpainting with iterative confidence feedback and guided upsampling. In *ECCV*, 2020. 1, 2, 3

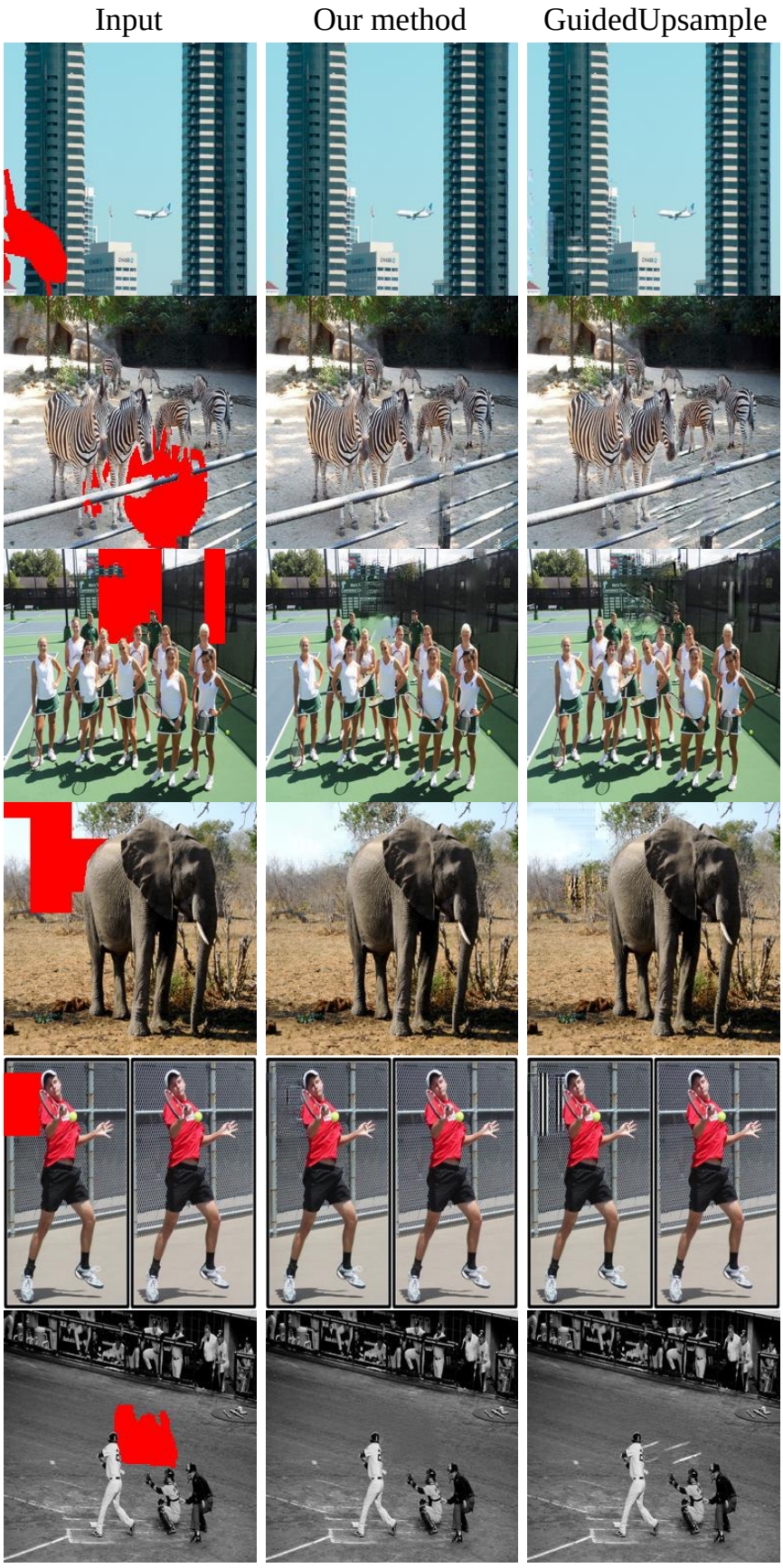

Figure 2. Qualitative comparison between our method and guided upsampling model in [2].

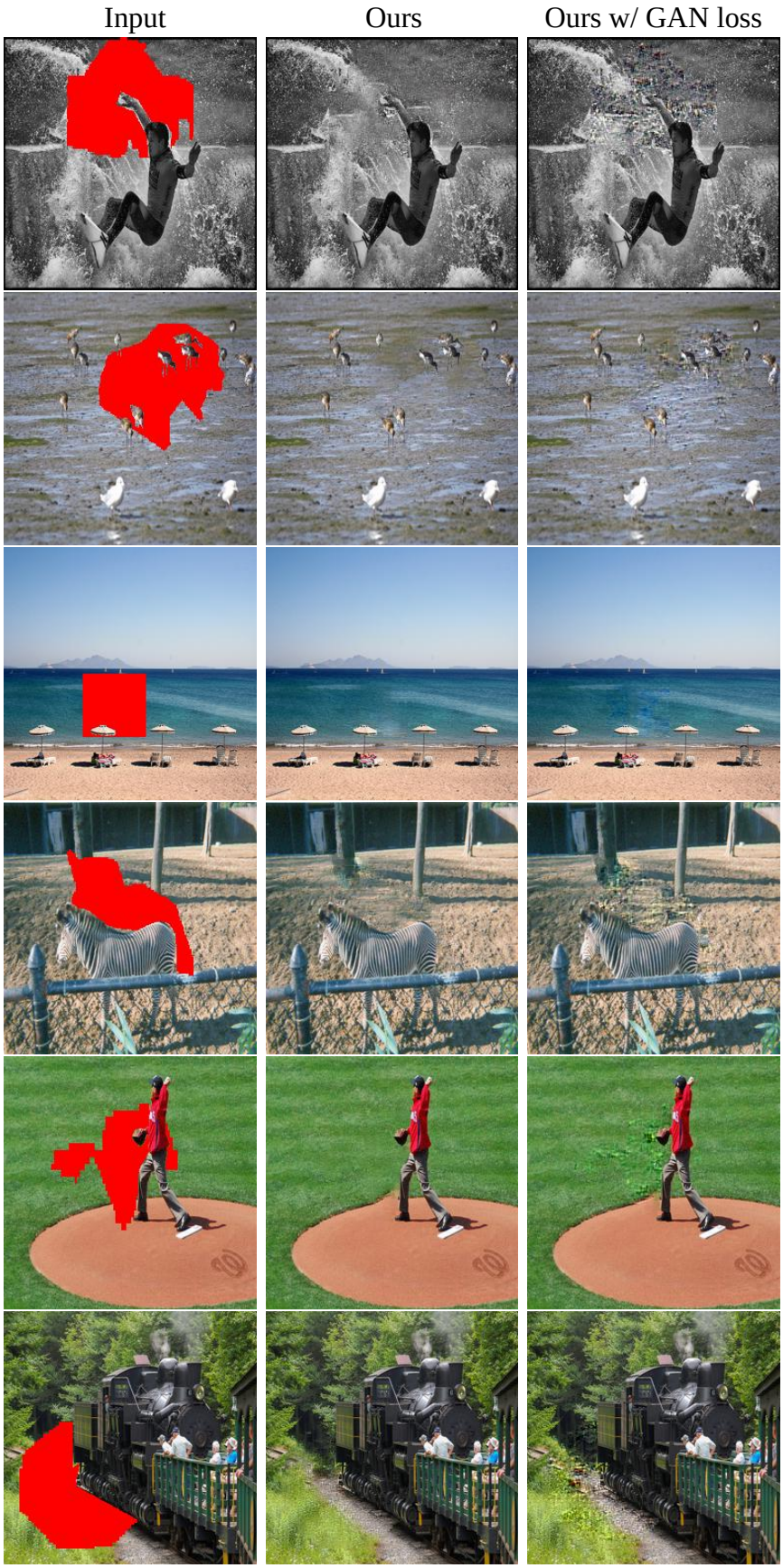

Figure 3. Qualitative comparison between our method trained without GAN loss and that trained with GAN loss.

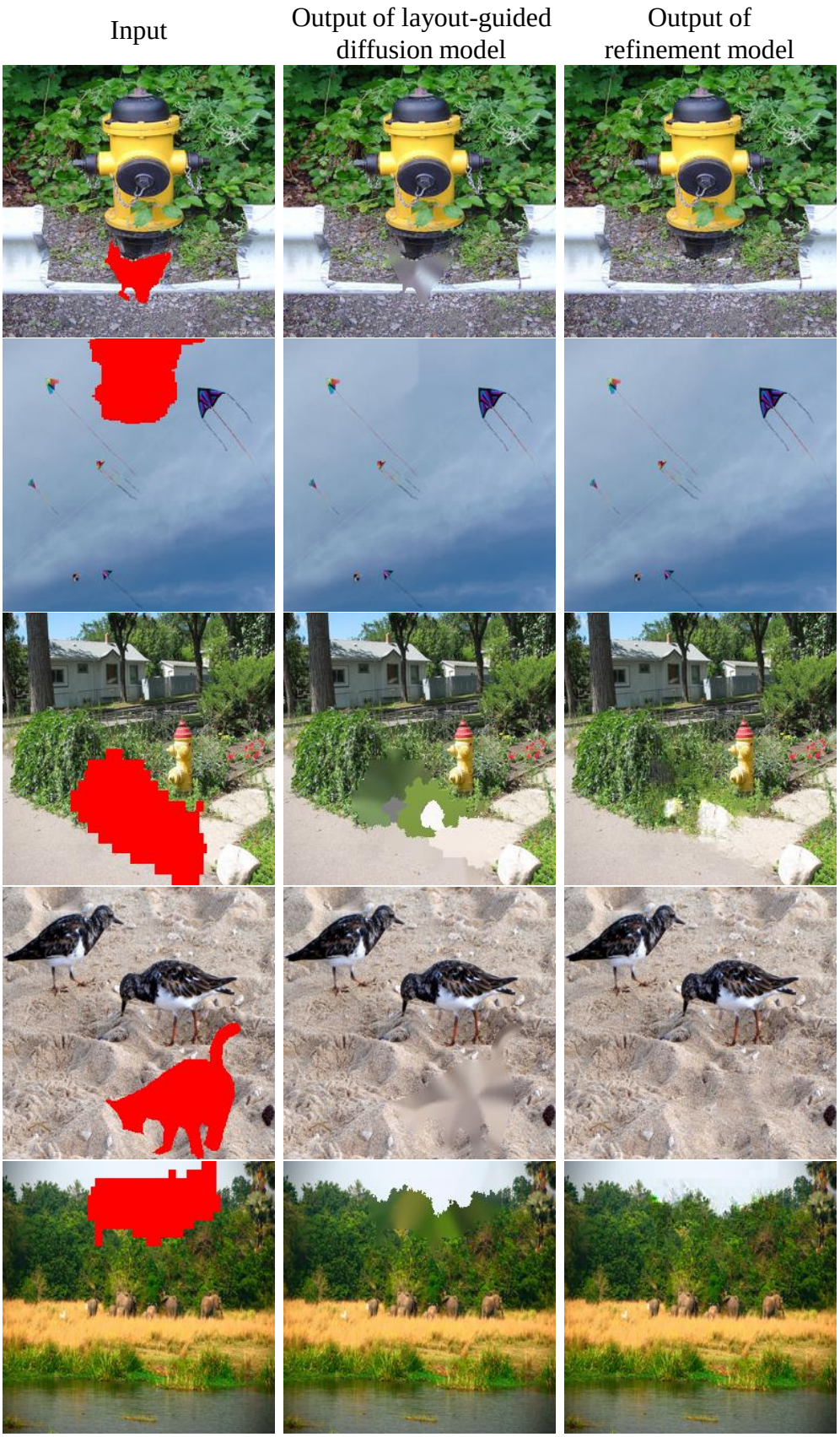

Figure 4. Output of each sub-model in our method.

Figure 5. “Contextual attention” vs. “Top-X attention” in effects. Supplement to Fig. 4 (b) of the paper.

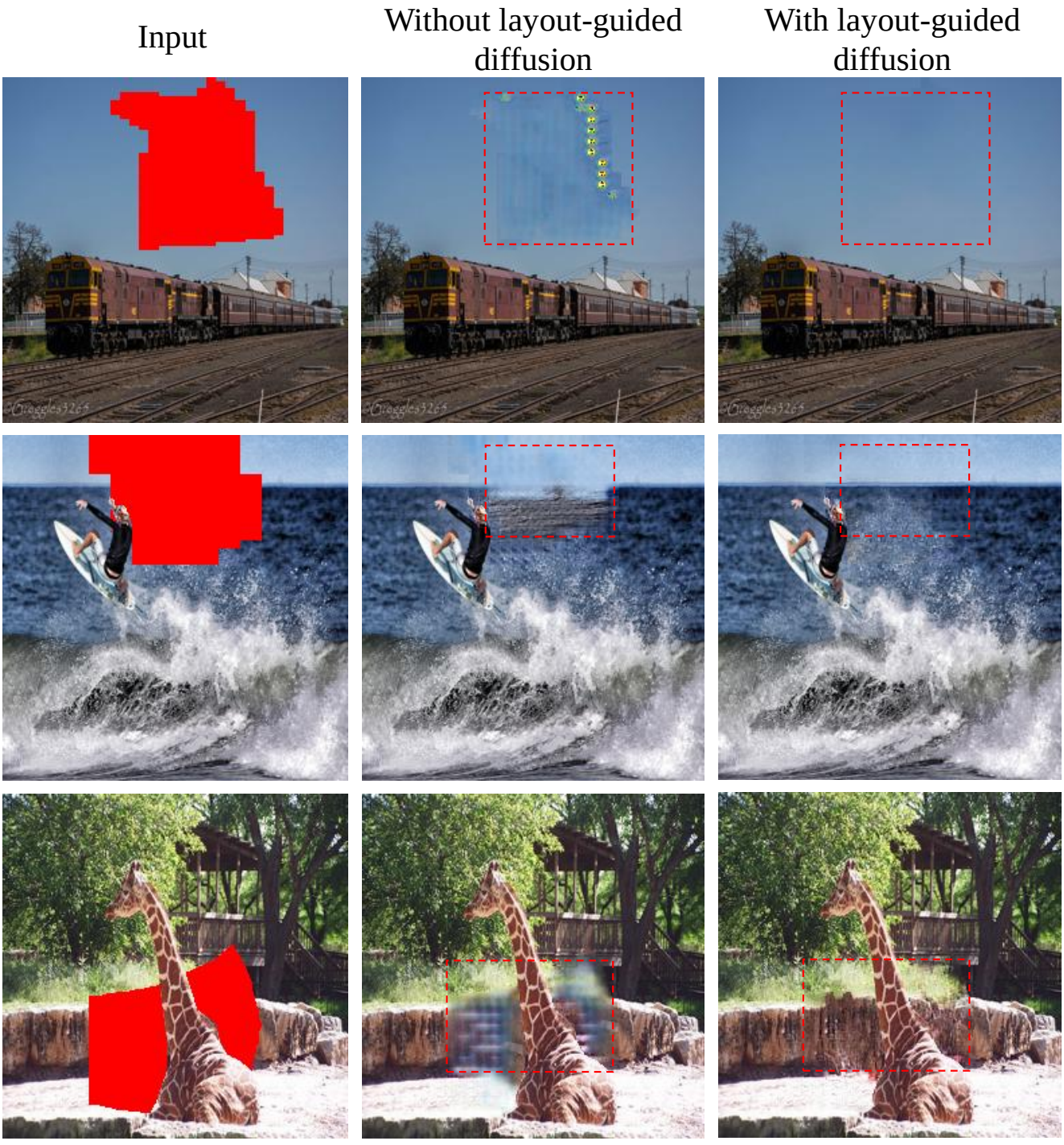

Figure 6. Without/with layout-guided diffusion method in effects.

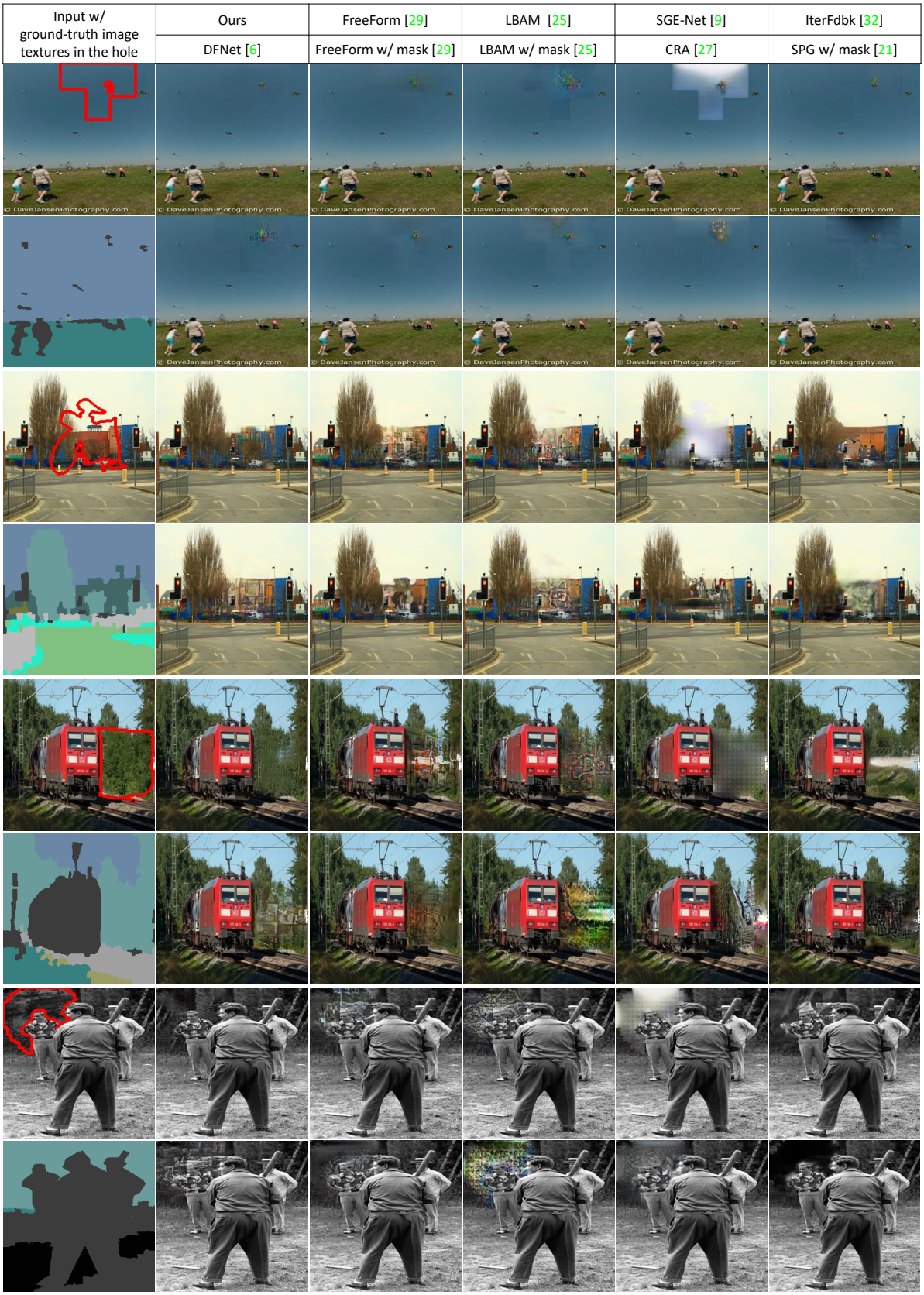

Figure 7. Qualitative comparison in mixed scenes. The top table indicates the display order. Textures enclosed by the red contours in the first column are to be restored. Supplement to Fig. 5 of the paper. The citation indices correspond to the references in the paper.

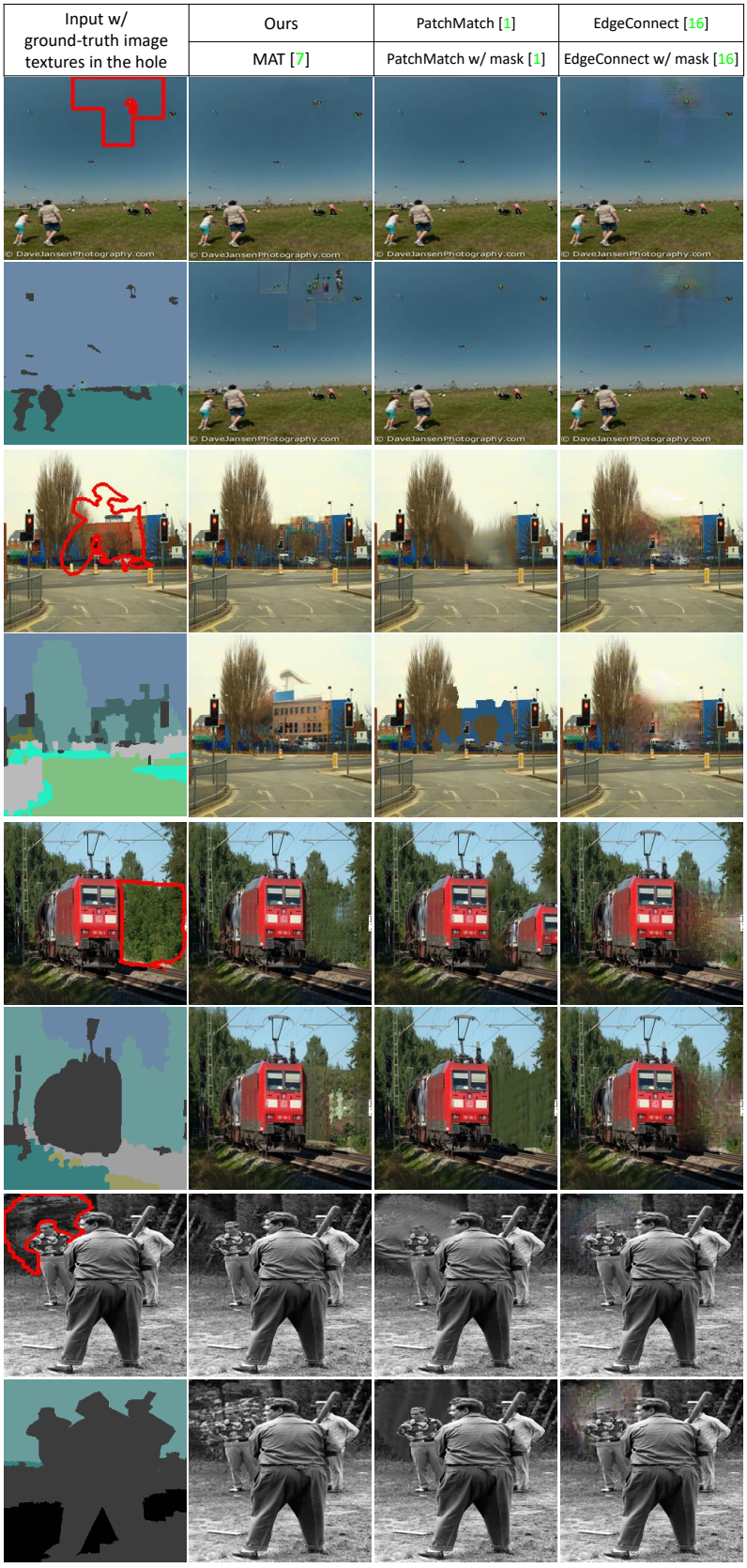

Figure 8. Qualitative comparison in mixed scenes. The top table indicates the display order. Textures enclosed by the red contours in the first column are to be restored. Supplement to Fig. 5 of the paper. The citation indices correspond to the references in the paper.

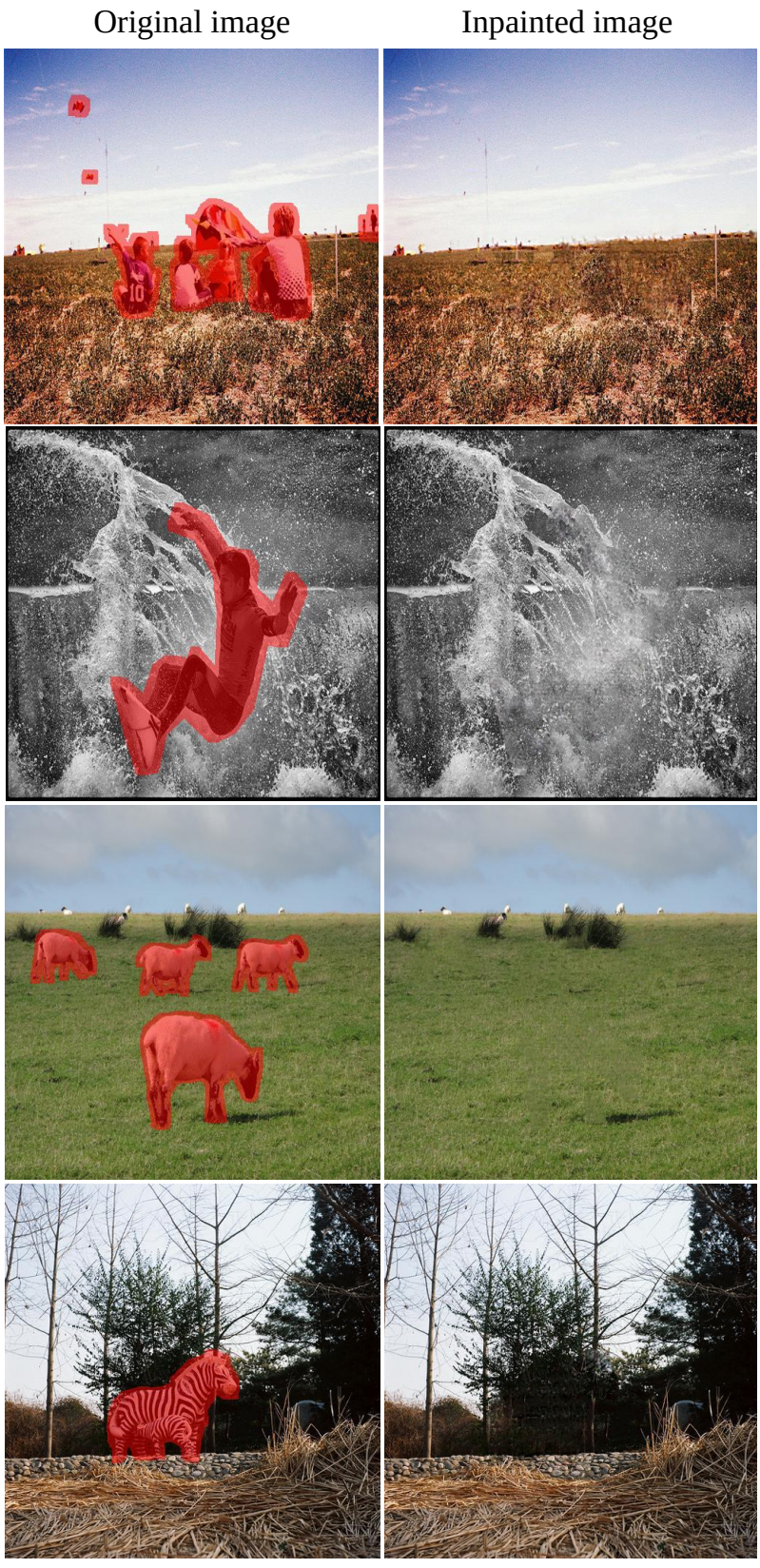

Figure 9. High-resolution (512×512) object removal.

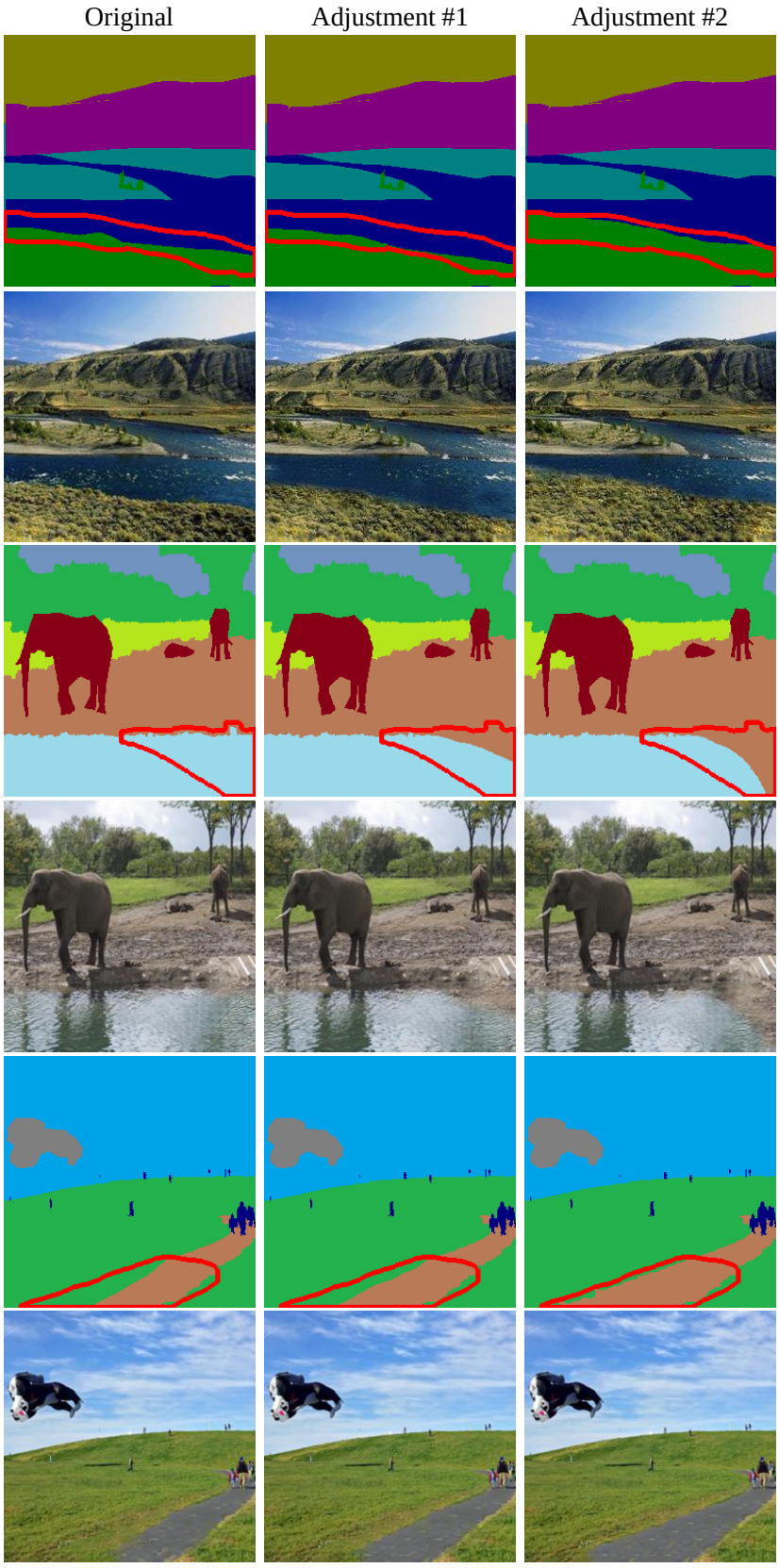

Figure 10. Object contour adjustment. The contour of objects in the image can be adjusted accordingly with the layout mask.

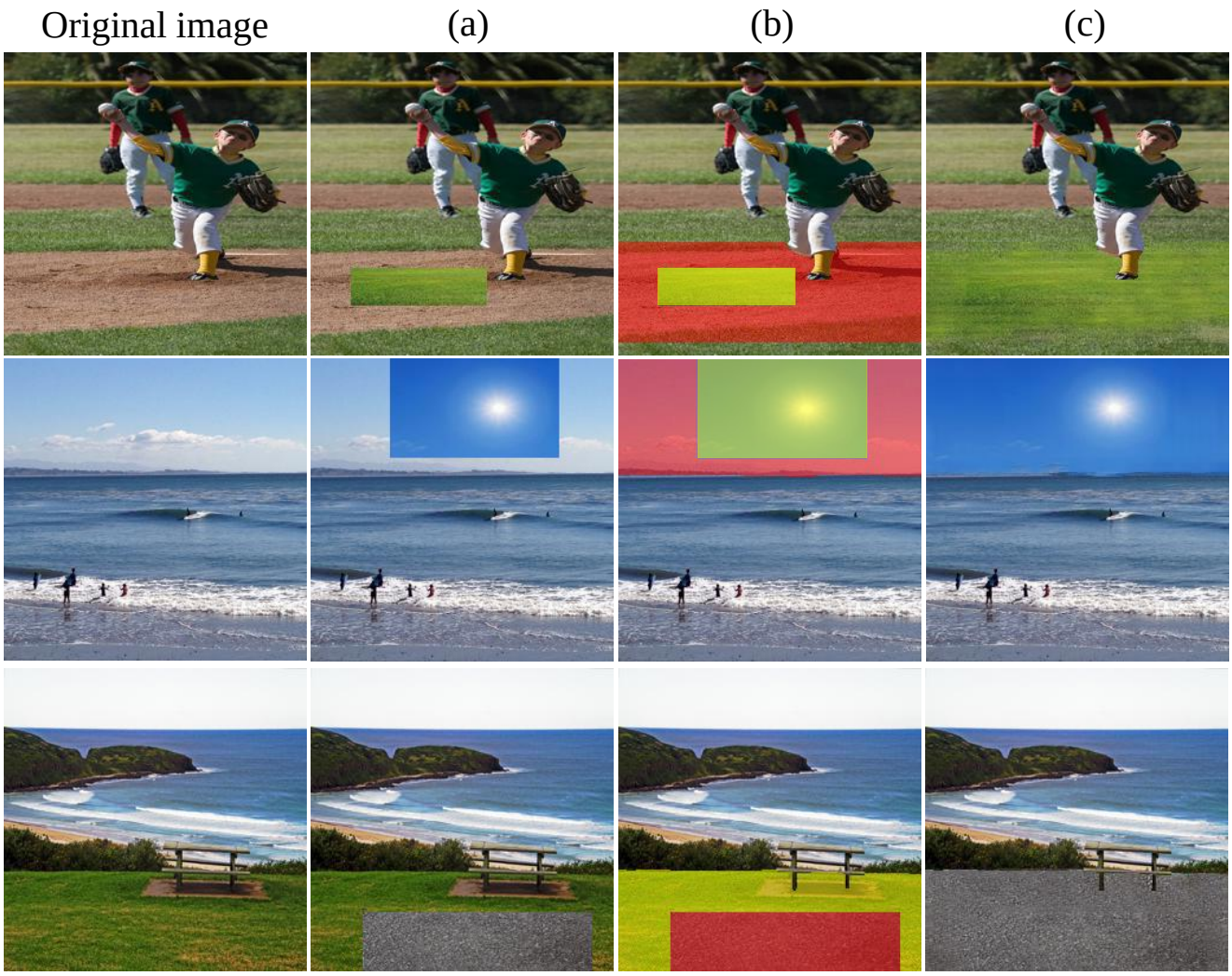

Figure 11. Template based content creation. The input image *(a)* is generated by pasting a new texture template on the original image. The pasted area and its relevant contextual area are denoted by the overlaid red and yellow masks in *(b)*. Our method generates content for the red area (*viz.*, hole regions) based on the pasted area (*viz.*, non-hole regions) and the context relevance. The generation results are shown in *(c)*.