Supplemetary Material:
Interactive Image Synthesis with Panoptic Layout Generation

Bo Wang Tao Wu Minfeng Zhu Peng Du
Huawei Technologies
{wangbo341, zhuminfeng, dupeng25}@hisilicon.com, taowu1@huawei.com

A. Supplementary Results

A.1. Qualitative Results on Visual Genome

In Fig. 8, we show visual comparisons of LostGAN-V1 [4], CAL2I [2], and the proposed PLGAN using perturbed Bounding Boxes as input based on the VG dataset [3]. In Fig. 9, we show sythesized image samples with the corresponding panoptic layouts on the VG dataset under 128^2 and 256^2 resolutions.

A.2. Quantitative Results on Visual Genome

Similar to Tab. 1 in the main paper, Tab. 4 reports quantitative comparison with respect to Inception Score, FID and CAS on the VG dataset.

A.3. Qualitative Results on Landscape

In Fig. 10, we compare generated images from Grid2Im [1] and our PLGAN on the Landscape dataset.

A.4. Quantitative Results on Landscape

In Tab. 5, we quantitavely compare Grid2Im [1] and our PLGAN on the Landscape dataset, for which all objects are "stuff". Our method outperforms Grid2Im on all metrics. The fact that this dataset contains only stuff objects makes the difference even more apparent.

A.5. Robustness to Perturbed BBoxes

In Fig. 11, we plot IS, FID and Coverage curves with varying perturbation range for Grid2Im [1], LostGAN-V2 [5] and our PLGAN under 256^2 resolution. Similar to the robustness test under 128^2 resolution in Fig. 7, PLGAN again claims the most robust model among others.

A.6. User Study

We conduct a user study on Wjx (https://www.wjx.cn) to rate realism of generated images. Specifically, we select 100 grouped image samples generated from Grid2Im, LostGAN, CAL2I and our method under 128×128 resolution. Each vote picks one of the two images from the same group and counts one point for the corresponding image generator. The overall scores after 600 votes in total are shown in Tab. 6.

B. Guide Filter

In Figure 12, we illustrate the workflow of the Guided Filter module. First, a 3×3 convolution layer is used to map the image feature X to tensor X_g of three channels. Following DGF [6], each instance layout L_i^{Th} and X_g are filtered by a prescribed 3×3 convolution kernel. And the linear transformation parameters $A \in \mathbb{R}^{H \times W \times 1}$ and $b \in \mathbb{R}^{H \times W \times 1}$ are predicted from CNN layers. Specifically, mean filter and covariance operations are carried out sequencially to get $\overline{X}_g$, $\overline{L}_i^{Th}$, $\Sigma_{\overline{X}_g, \overline{X}_g}$ and $\Sigma_{\overline{X}_g, \overline{L}_i^{Th}}$. Then the parameter A is predicted from $\Sigma_{\overline{X}_g, \overline{X}_g}$ and $\Sigma_{\overline{X}_g, \overline{L}_i^{Th}}$ by Convolution Block, which contains 3 conditional layers with 1×1 kernels. And the parameter b is equal to $\overline{L}_i^{Th} - A \odot \overline{X}_g$. Finally, the refined layout is computed as follows:

$$\widetilde{L}_i^{Th} = A \odot L_i^{Th} + b. \tag{1}$$

References

[1] Oron Ashual and Lior Wolf. Specifying object attributes and relations in interactive scene generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4561–4569, 2019. 1, 4

[2] Sen He, Wentong Liao, Michael Yang, Yongxin Yang, Yi-Zhe Song, Bodo Rosenhahn, and Tao Xiang. Context-aware layout to image generation with enhanced object appearance. In *CVPR*, 2021. 1, 4

[3] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 1

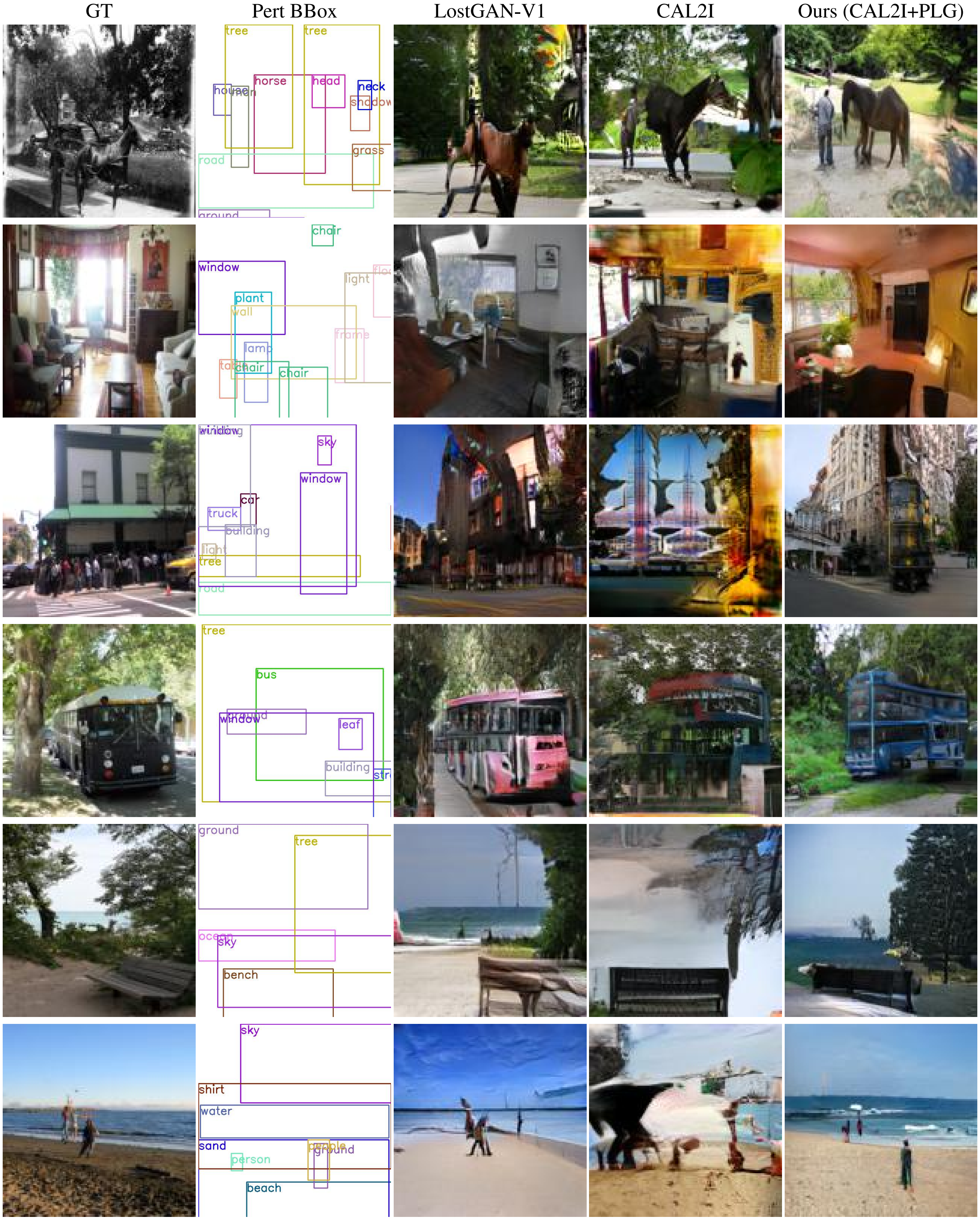

Figure 8. Visual comparison between sample images generated from perturbed BBoxes (Pert BBoxes) on the VG dataset.

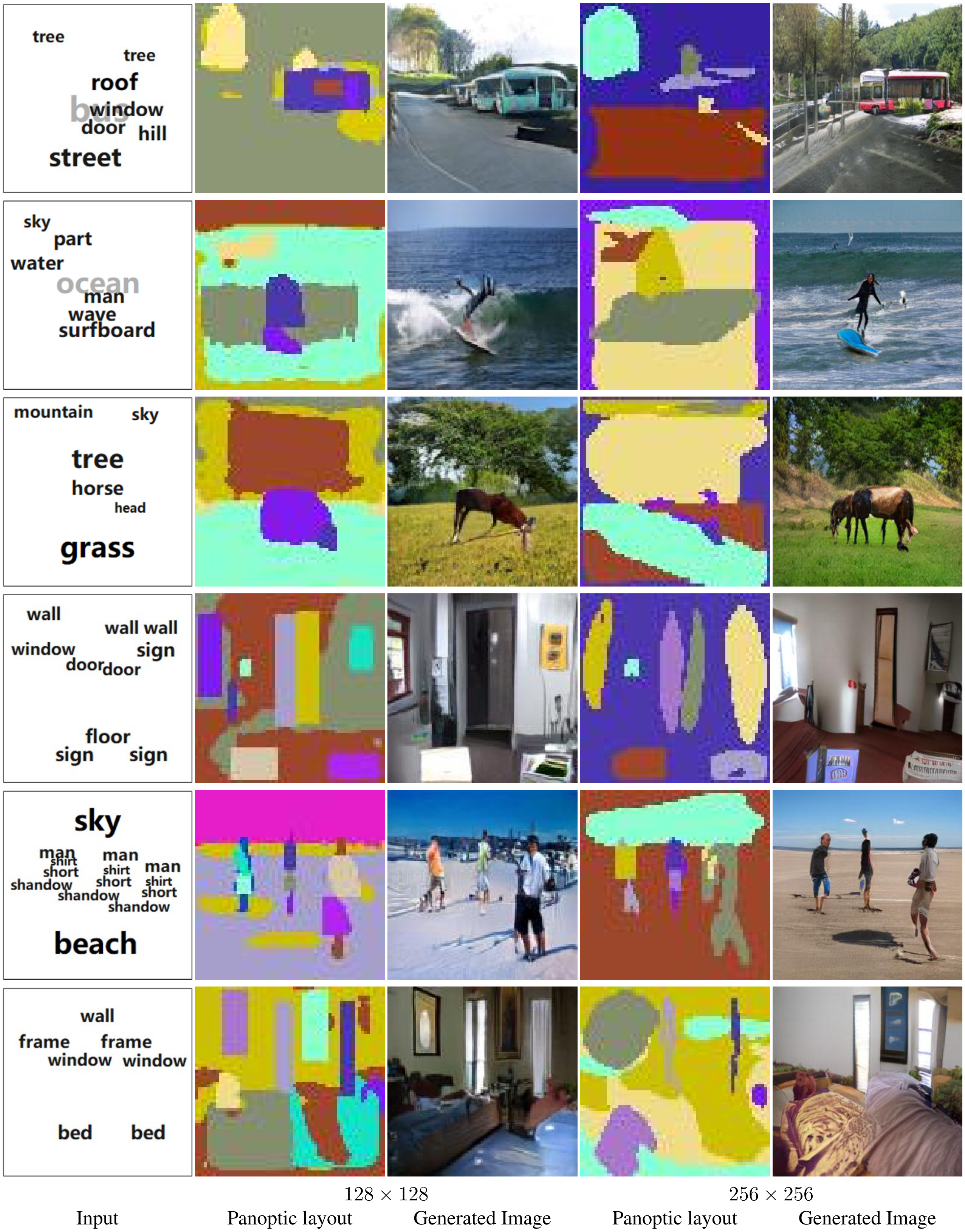

Figure 9. Synthesized image samples on the VG dataset.

Table 4. Quantitative comparison with respect to Inception Score, FID and CAS on the VG dataset.

Methods	Resolution	IS↑			FID↓			CAS↑
Real Images	128×128	20.5±1.5			-			48.07
Real Images	256×256	28.6±1.1			-			
		GT BBox	Pert1 BBox	Pert2 BBox	GT BBox	Pert1 BBox	Pert2 BBox	GT BBox
LostGAN-V1 [4]	128×128	11.1±0.6	10.3±0.1	9.7±0.1	29.36	39.48	42.29	28.85
LostGAN-V2 [5]		10.7±0.2	-	-	29.00	-	-	29.35
CAL2I [2]		12.6±0.4	8.4±0.1	7.3±0.1	21.78	49.53	61.30	29.2
Ours(CAL2I [2]+PLG)		**12.7±0.2**	**10.6±0.1**	**10.1±0.1**	**20.62**	**32.93**	**37.03**	**30.81**
LostGAN-V2 [5]	256×256	14.1±0.3	-	-	47.62	-	-	28.81
Ours(CAL2I [2]+PLG)		**14.9±0.1**	13.2±0.2	12.6±0.1	**28.06**	38.41	41.36	**29.35**

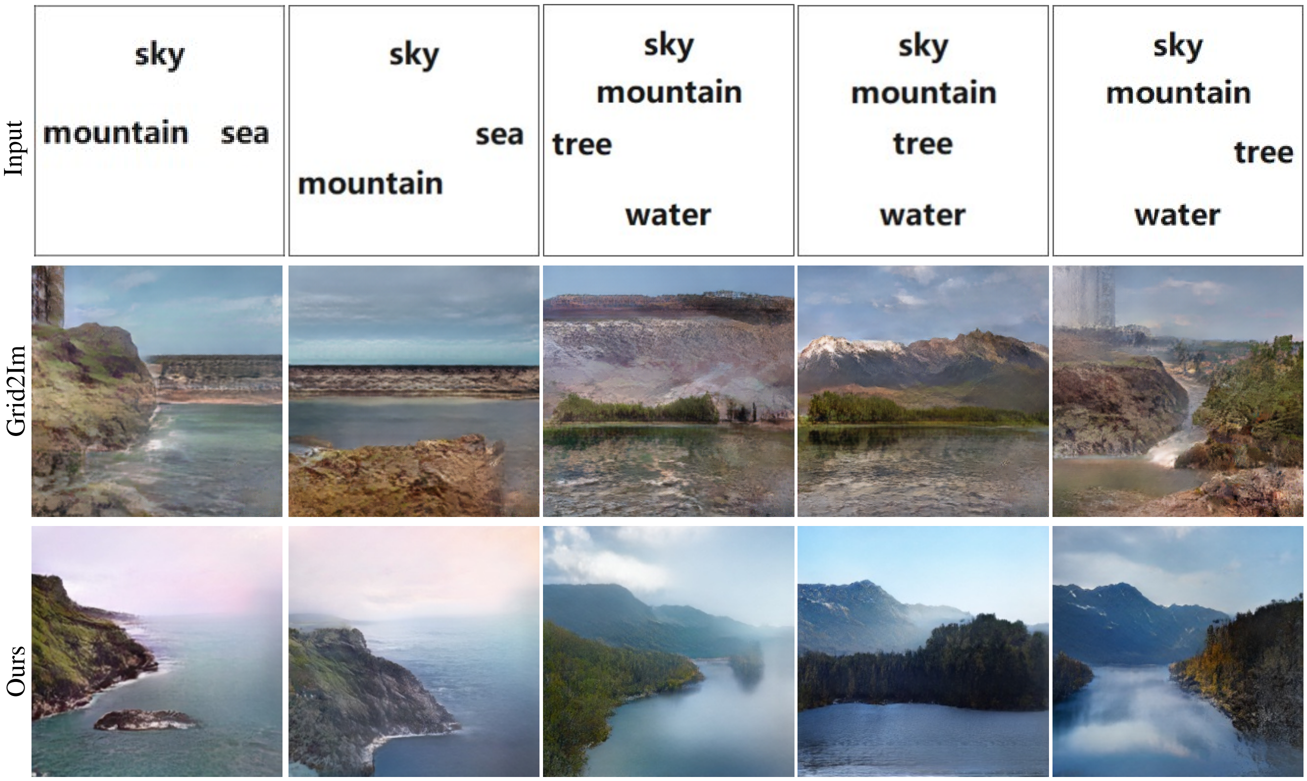

Figure 10. Visual comparison on the Landscape dataset.

Table 5. Quantitative comparison on the Landscape dataset.

Method	Resolution	IS↑	FID ↓
Real Images	448×448	5.9±0.2	-
Grid2Im [1]		1.8±0.1	144.84
Ours		3.3±0.1	57.40

Table 6. User Study statistical results.

Grid2Im	LostGAN	CAL2I	PLGAN(ours)
118	142	143	197

[4] Wei Sun and Tianfu Wu. Image synthesis from reconfigurable layout and style. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10531–10540, 2019.

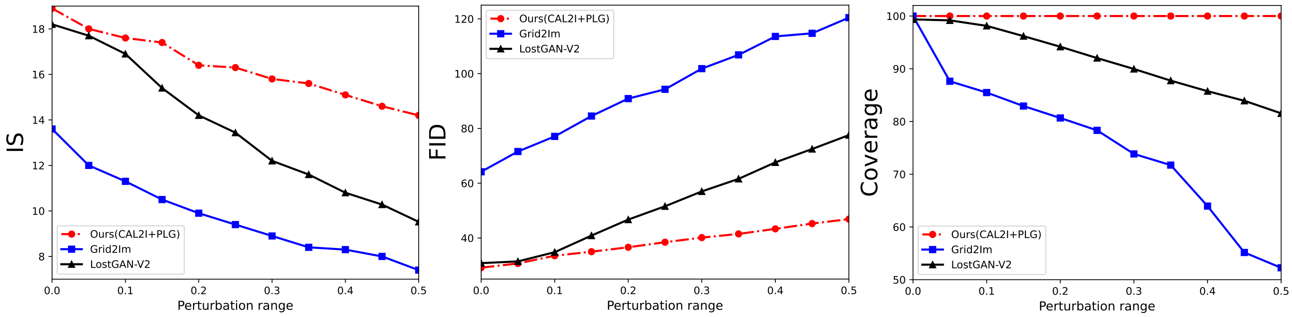

Figure 11. IS, FID and Coverage curves with varying perturbation range on the COCO-Stuff dataset under 256×256 resolution.

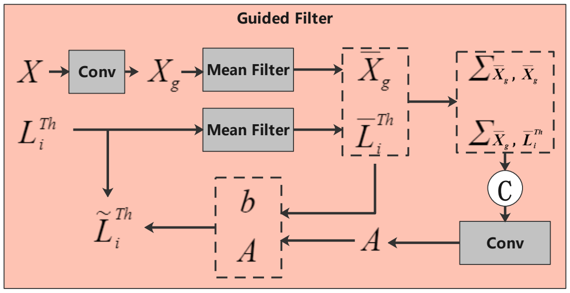

Figure 12. Workflow of Guided Filter.

1, 4

[5] Wei Sun and Tianfu Wu. Learning layout and style reconfigurable gans for controllable image synthesis. *arXiv preprint arXiv:2003.11571*, 2020. 1, 4

[6] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Fast end-to-end trainable guided filter. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1838–1847, 2018. 1