PLACE: Adaptive Layout-Semantic Fusion for Semantic Image Synthesis

Supplementary Material

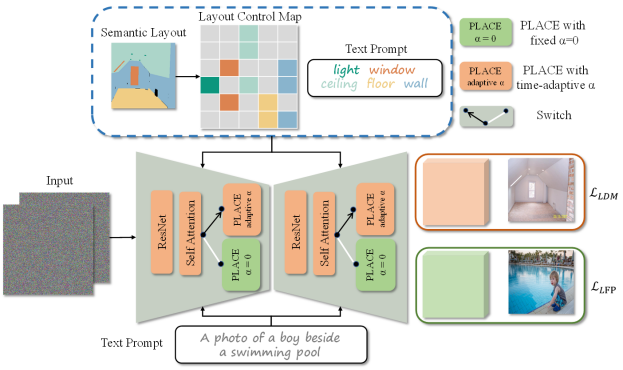

Figure A. Fine-tuning with Layout-Free Prior Preservation Loss.

This supplemental file provides the following materials:

- More details of LFP Loss in Sec. A;
- More ablation studies in Sec. B;
- More qualitative results in Sec. C;
- Discussion on limitation in Sec. D.

A. More Details of LFP Loss

During each fine-tuning iteration, we sample a set of image-semantic layout pairs $< z_t, S, y, t >$ and a set of image-text pairs $< z'_t, y', t' >$ from the Layout Free dataset. As depicted in Fig. A, for training data annotated with semantic layouts, we employ the PLACE with timestep-adaptive α to synthesize images and compute the semantic image synthesis loss $\mathcal{L}_{LDM}$. For image-text training data, we adopt the PLACE with fixed $\alpha(\alpha = 0)$ to synthesize images and calculate the layout-free prior preservation loss $\mathcal{L}_{LFP}$.

B. More Ablation Study Results

B.1. Layout Control Map

Fig. B illustrates the impact of the Layout Control Map (LCM) on the generated images. The $2nd$ and $3rd$ columns, as well as the $4th$ and $5th$ columns, represent layout presentations (64×64) without LCM and with the utilization of LCM, respectively, along with their corresponding synthesized images. It can be observed that LCM preserves more details in the low-resolution feature space, thus promoting faithful details and improved layout consistency.

B.2. Adaptive α for fusion

Fig. C displays the synthesis results with and without Layout-Semantic Adaptive Fusion. The $2nd$ and $4th$ columns represent the fusion maps corresponding to fixed alpha and adaptive alpha, respectively. It can be seen that adaptive fusion preserves the interactions of specified semantics (such as 'stairs' or 'car') on relevant semantic regions (such as 'house' or 'road'), resulting in the synthesis of more realistic details and higher visual quality.

Figure B. More Qualitative Ablation Comparisons on LCM. The $2nd$ and $4nd$ depict the layout representation for part of semantics.

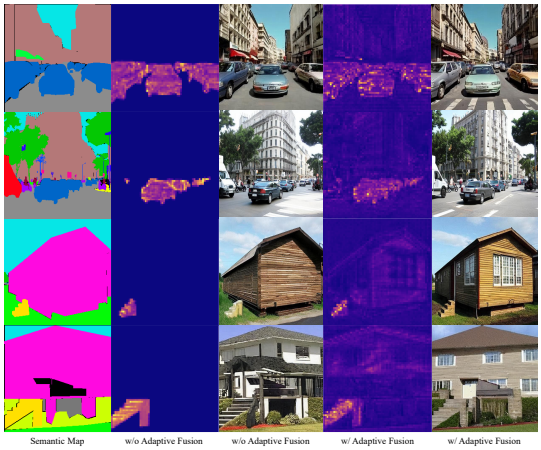

Figure C. More Qualitative Comparisons on Adaptive Fusion. The $2nd$ and $4nd$ show the fusion map for part of semantics.

B.3. Semantic Alignment loss

Fig. D presents the results of the model fine-tuned with and without Semantic Alignment loss. The $2nd$ and $4th$ columns respectively show the self-attention maps of models fine-tuned without and with SA loss. It can be observed that the SA loss facilitates the interaction of image tokens

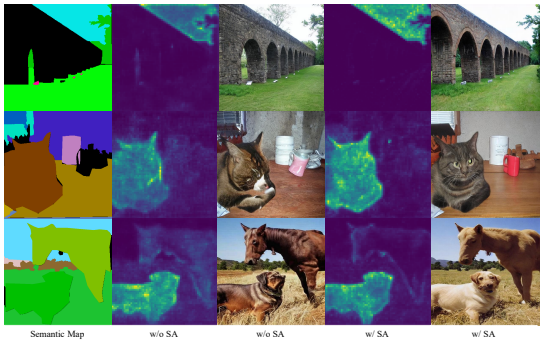

Figure D. More Qualitative Comparisons on SA Loss. The $2nd$ and $4nd$ show the self-attention map for part of semantics.

within the same or related semantic regions (*e.g.*'sky', 'cat', and 'dog'), thereby improving the layout consistency and visual quality of synthesized images.

B.4. Layout Free Prior Preservation loss

In this section, we first present more ablation qualitative comparison results (new object synthesis) on the Layout Free Prior Preservation loss in Fig. E. Additionally, we validate the effectiveness of our layout-free prior preservation loss by assessing the original text-to-image synthesis capability of different models, as shown in Table A and Fig. F.

Firstly, as observed from Fig. E, the utilization of LFP Loss results in enhanced visual quality in the synthesis of semantic images. Notably, 'cat' in the $2nd$ row, 'sheep' in the $3rd$ row, and 'laptop' in the $6th$ row, all demonstrate improved visual results. Additionally, the semantic consistency of the results has been elevated, as shown in the $4th$ row with 'fog' and the $5th$ row with 'clouds'. These results collectively substantiate the effectiveness of the LFP Loss.

Then we assessed the original text-to-image synthesis capabilities of four models: Original Stable Diffusion V1-4 (SD V1-4), FreestyleNet, our model without using LFP (Ours w/o LFP), and our model with LFP (Ours w/ LFP). In this case, both our models and FreestyleNet were fine-tuned on the ADE20K dataset using SD V1-4 as the initial parameters. We extract 1500 captions as input text prompts from the validation set of COCO-Stuff. During sampling, we employ 50 PLMS sampling steps with a classifier-free guidance scale of 2. We calculate the FID and Text-Alignment scores between the synthesized images and ground truth, as shown in Table A. It can be observed that the adoption of the LFP Loss significantly preserves the original text-to-image synthesis capability of the fine-tuned model. The FID decreases from 46.2 to 36.8, and Text-Alignment increases from 0.27 to 0.30, approaching the performance of the original model. This indicates the important role of LFP loss in preserving semantic concepts in the original model. Furthermore, the qualitative comparisons in Fig. F also indicate our LFP Loss helps preserve priors in the original model, resulting in the synthesis of images with improved semantic consistency and visual quality.

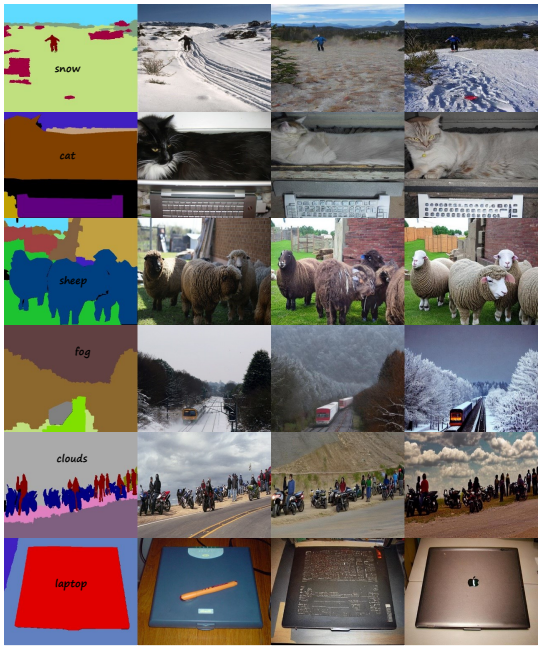

Figure E. More Qualitative Ablation Comparisons on LFP Loss.

Method	FID↓	Text-Alignment ↑
SD V1-4	**34.4**	**0.31**
FreestyleNet	47.4	0.27
Ours w/o LFP	46.2	0.27
Ours w/ LFP	<u>36.8</u>	<u>0.30</u>

Table A. Quantitative comparison of Text-to-Image Synthesis.

C. More Qualitative Results

C.1. Additional Details of OOD Evaluation

For quantitative evaluation of Out-Of-Distribution(OOD) synthesis, we conduct experiments from three aspects: new object, new attribute, and new style. We assess the generalization capability of the model fine-tuned on ADE20K. For new object synthesis, we utilize the model to synthesize 5000 images in the validation set of COCO-Stuff. We compute the FID score of synthesized images and the mIoU of semantic classes exclusive to COCO-Stuff. For new attribute synthesis, a total of 260 images are synthesized for six attributes: "*brick* wall", "sky *with rainbow*", "*autumn* flora/tree/grass", "*wooden* floor", "*snowy* road", and "*colorful* carpet". The text alignment between images and text prompts is computed with CLIP. For new style synthesis, a total of 260 images are synthesized across eight different styles: "drawn by Van Gogh", "in oil painting", "in Minecraft", "full of graffiti", "in sketch", "in Monet style", "in anime" and "drawn by Picasso." The text alignment between images and text prompts is assessed by CLIP. Additionally, the semantic layouts used for evaluating new attribute and new style are both sampled from the ADE20K.

Figure F. Comparison of Original Text-to-Image Synthesis results across different models.

C.2. Out-of-distribution Synthesis.

We show more visual comparisons with the competing methods on out-of-distribution synthesis. Fig. G, Fig. H and Fig. I respectively present the qualitative comparisons of the new object, new attribute, and new style, which are synthesized by a model fine-tuned on the ADE20K dataset. It can be observed that our approach synthesizes results that are more consistent with the provided text input.

C.3. In-distribution Synthesis.

Fig. J and Fig. K respectively illustrate more in-distribution qualitative comparisons on the ADE20K and COCO-Stuff dataset. It can be seen that our synthesized images are not only of high fidelity but also exhibit a strong alignment with the provided semantic layout in terms of finer details.

D. Limitation

Although PLACE has made advancements in visual quality, semantic consistency, and layout alignment, there are still some limitations. Firstly, the inference speed of diffusion-based methods is still slower compared to that of GAN-based methods. On a V100 GPU, ControlNet, FreestyleNet, and PLACE require an average of approximately 7.5 seconds (s), 5.9 s, and 6.1 s, respectively, to synthesize an image using PLMS sampling for 50 steps. We believe that with the development of superior samplers and latent consistency models, this issue will be largely alleviated. Additionally, constrained by the capabilities of the pre-trained stable diffusion, when prompts for a single class are too long or contain uncommon tokens, the synthesized image may be inconsistent with the given text. Higher-performance text-to-image models may potentially ameliorate this issue.

Figure G. Visual comparisons on new object synthesis.

Figure H. Visual comparisons on new attribute synthesis.

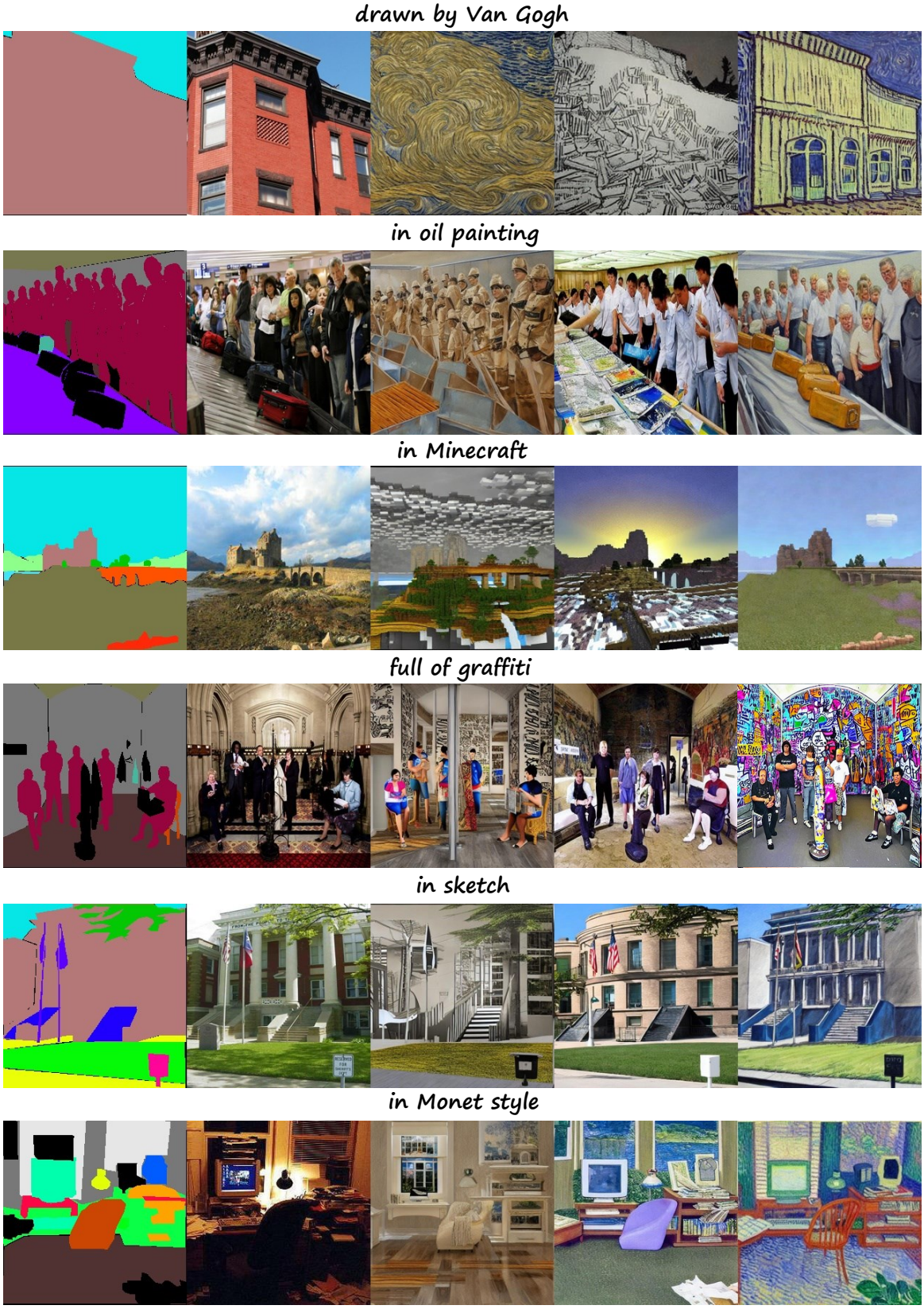

Figure I. Visual comparisons on new style synthesis.

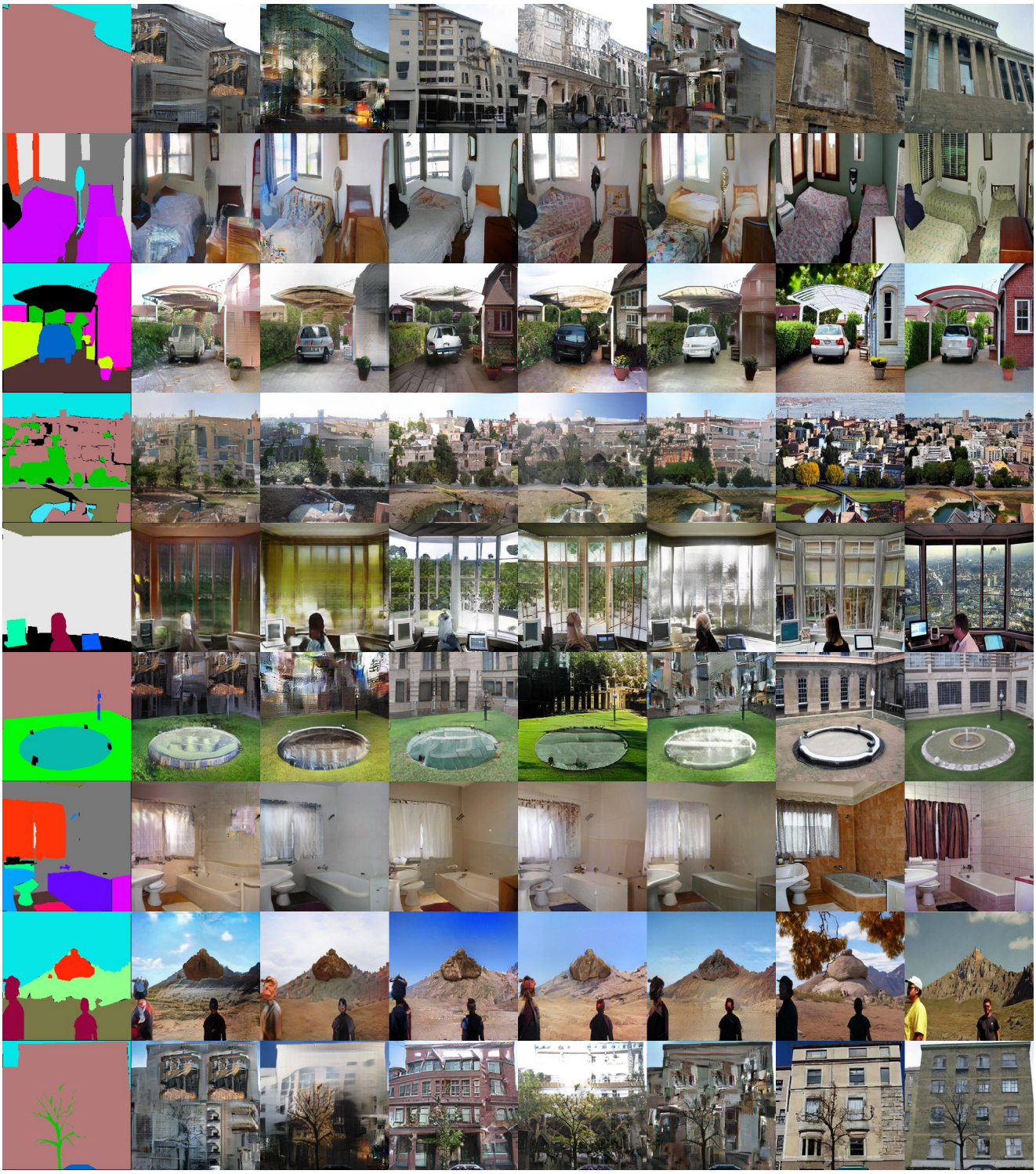

Figure J. Visual comparisons on ADE20K dataset.

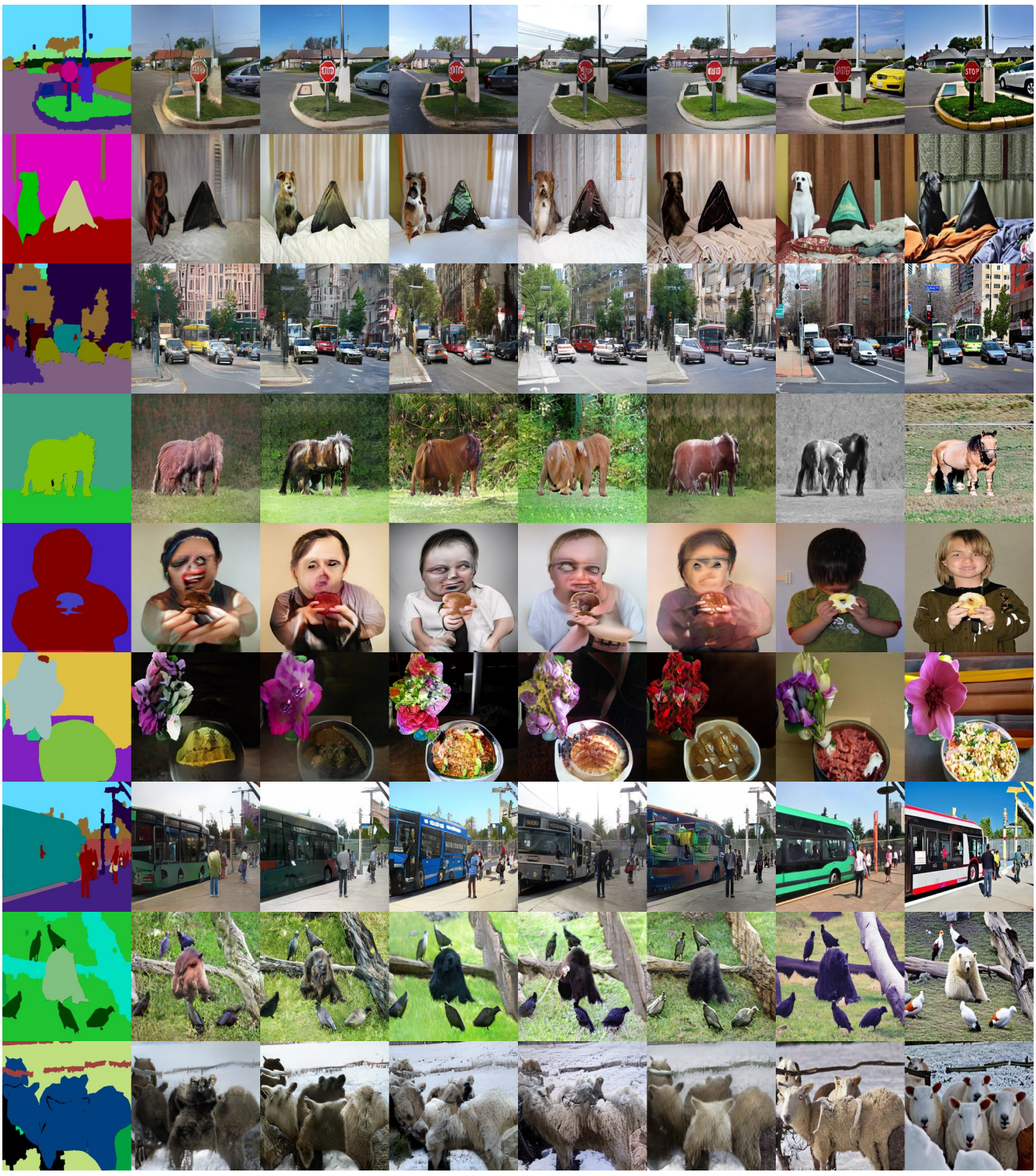

Figure K. Visual comparisons on COCO-Stuff dataset.