Supplementary Material *for* SCP-Diff: Photo-Realistic Semantic Image Synthesis with Spatial-Categorical Joint Prior

Huan-ang Gao*[1], Mingju Gao*[1], Jiaju Li[1,2], Wenyi Li[1],
Rong Zhi[3], Hao Tang[4], and Hao Zhao†[1]

[1] Institute for AI Industry Research (AIR), Tsinghua University
[2] University of Chinese Academy of Sciences
[3] Mercedes-Benz Group China Ltd.
[4] Peking University & Carnegie Mellon University
gha20@mails.tsinghua.edu.cn, gaomingju19@mails.ucas.ac.cn
zhaohao@air.tsinghua.edu.cn

1 Toolkit

For access to our fully anonymous code toolkit, please visit: https://anonymous.4open.science/r/SCP-Diff-Toolkit/.

2 Implementation Details

Finetuning ControlNet. We initialize the Stable Diffusion branch of ControlNet [8] with SD 2.1 weights. During training, we set the text prompt to fixed strings, and those are: *City road scenes* for Cityscapes [4], *Photorealistic and diverse images depicting various scenes* for ADE20K [9], and *high quality, detailed* for COCO-Stuff [1]. This aims to ensure that the text prompt remains devoid of any semantic cues, with the sole source of semantics derived from the semantic label processed by the control branch and the noise priors.

Evaluation. During inference of the generated results from different datasets, the text prompt was kept the same as the training procedure. For evaluation of FID, we sampled 50,000 images for each group of experimental setting, noting that FID score is biased and the bias is depending on the number of images we use for calculation [3]. For evaluation of mIoU, we follow OASIS [5], using UperNet101 [6] for ADE20K, multi-scale DRN-D-105 [7] for Cityscapes, and DeepLabV2 [2] for COCO-Stuff.

3 Results

3.1 More Qualitative Results

We provide more qualitative results, with Fig. 1 and Fig. 2 for Cityscapes, Fig. 3 and Fig. 4 for ADE20K, Fig. 5 and Fig. 6 for COCO-Stuff.

References

1. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1209–1218 (2018) 1
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062 (2014) 1
3. Chong, M.J., Forsyth, D.: Effectively unbiased fid and inception score and where to find them. arXiv preprint arXiv:1911.07023 (2019) 1
4. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016) 1
5. Sushko, V., Schönfeld, E., Zhang, D., Gall, J., Schiele, B., Khoreva, A.: Oasis: only adversarial supervision for semantic image synthesis. International Journal of Computer Vision **130**(12), 2903–2923 (2022) 1
6. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European conference on computer vision (ECCV). pp. 418–434 (2018) 1
7. Yu, F., Koltun, V., Funkhouser, T.: Dilated residual networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 472–480 (2017) 1
8. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023) 1
9. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017) 1

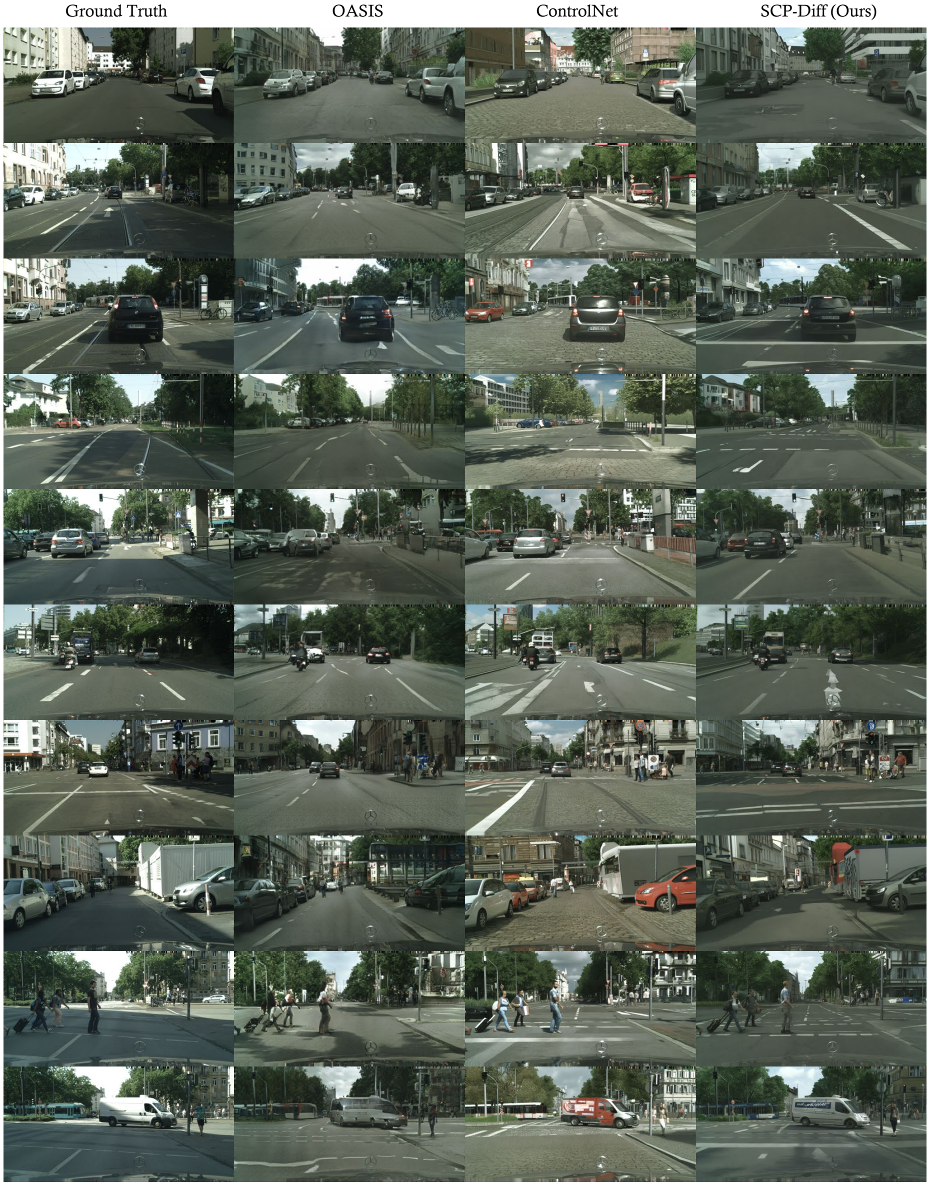

Fig. 1: Qualitative Results on Cityscapes dataset. (cont.)

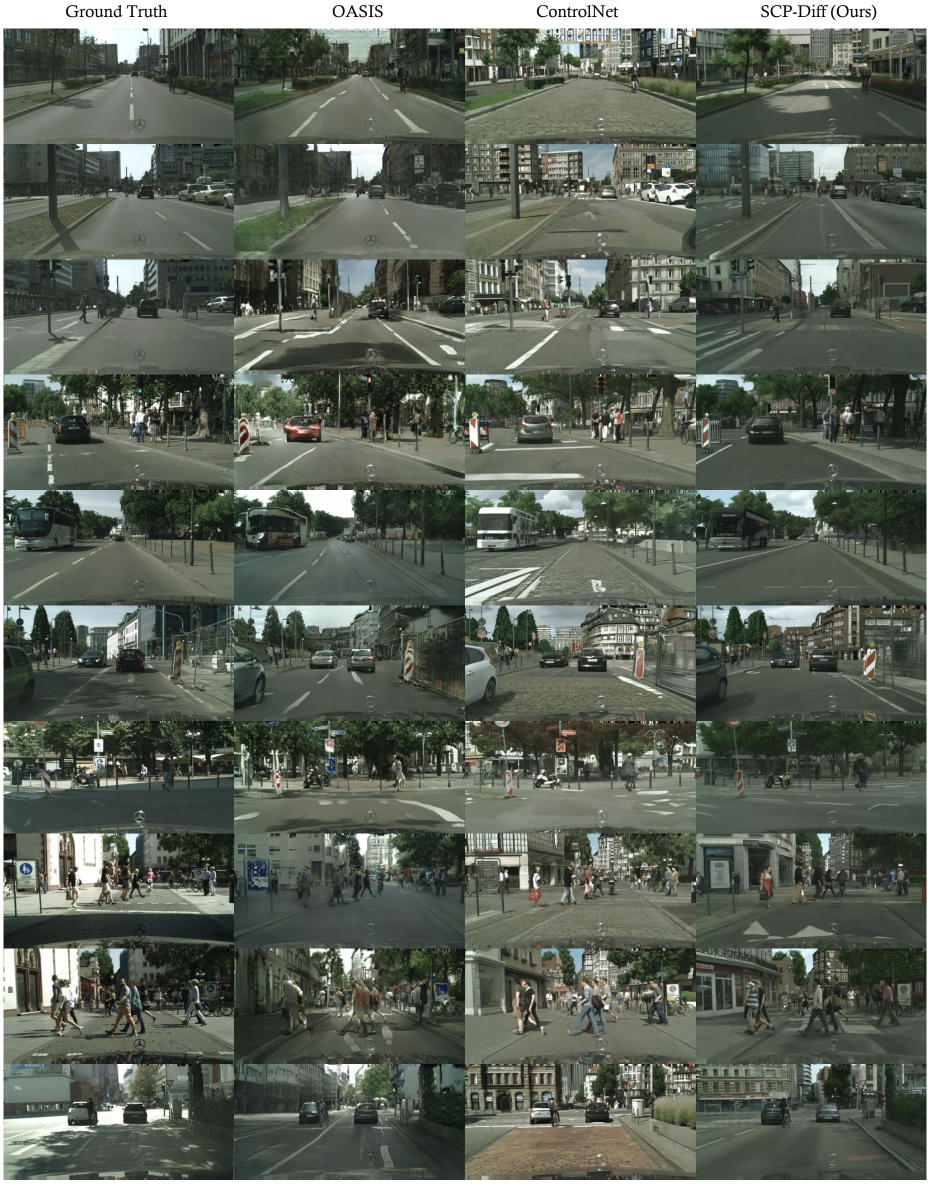

Fig. 2: Qualitative Results on Cityscapes dataset. (cont.)

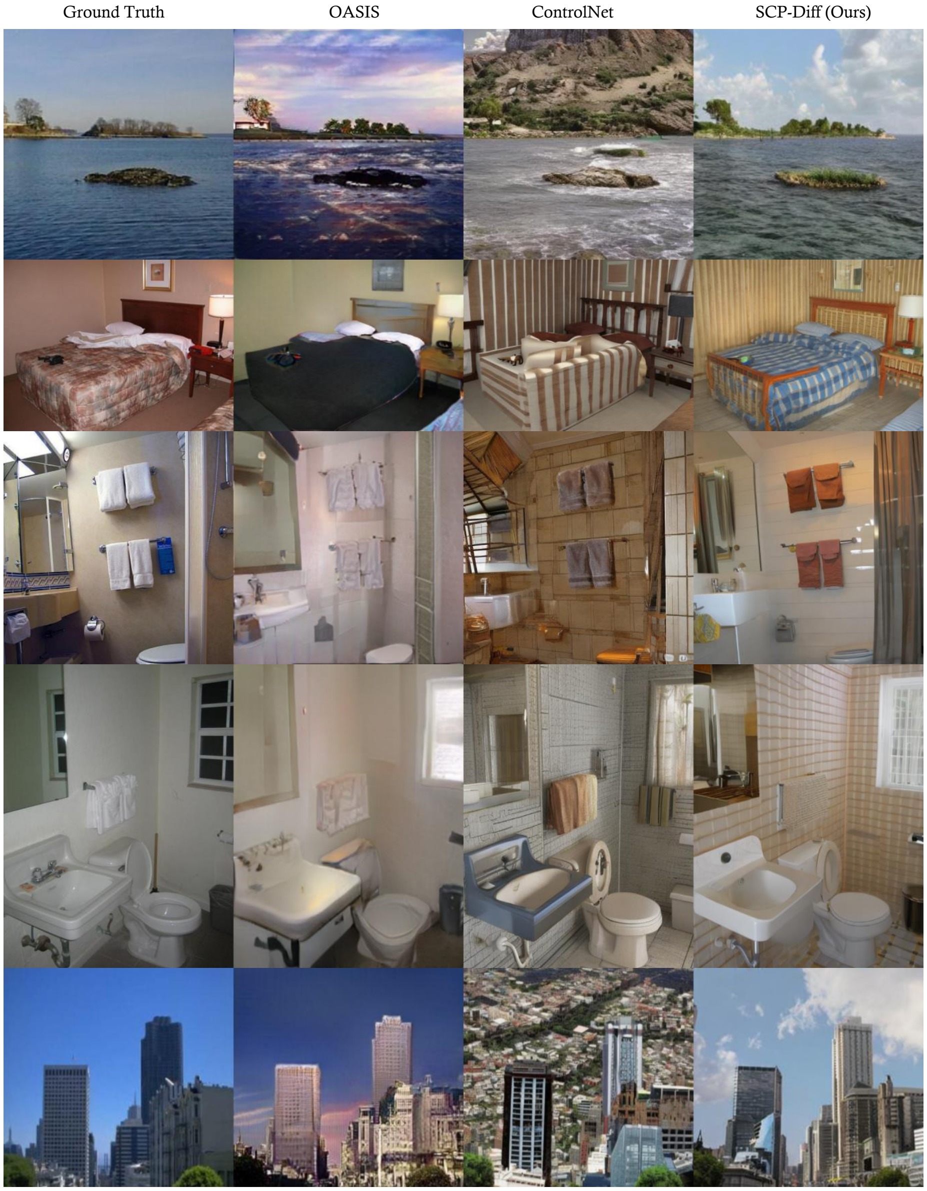

Fig. 3: Qualitative Results on ADE20K dataset. (cont.)

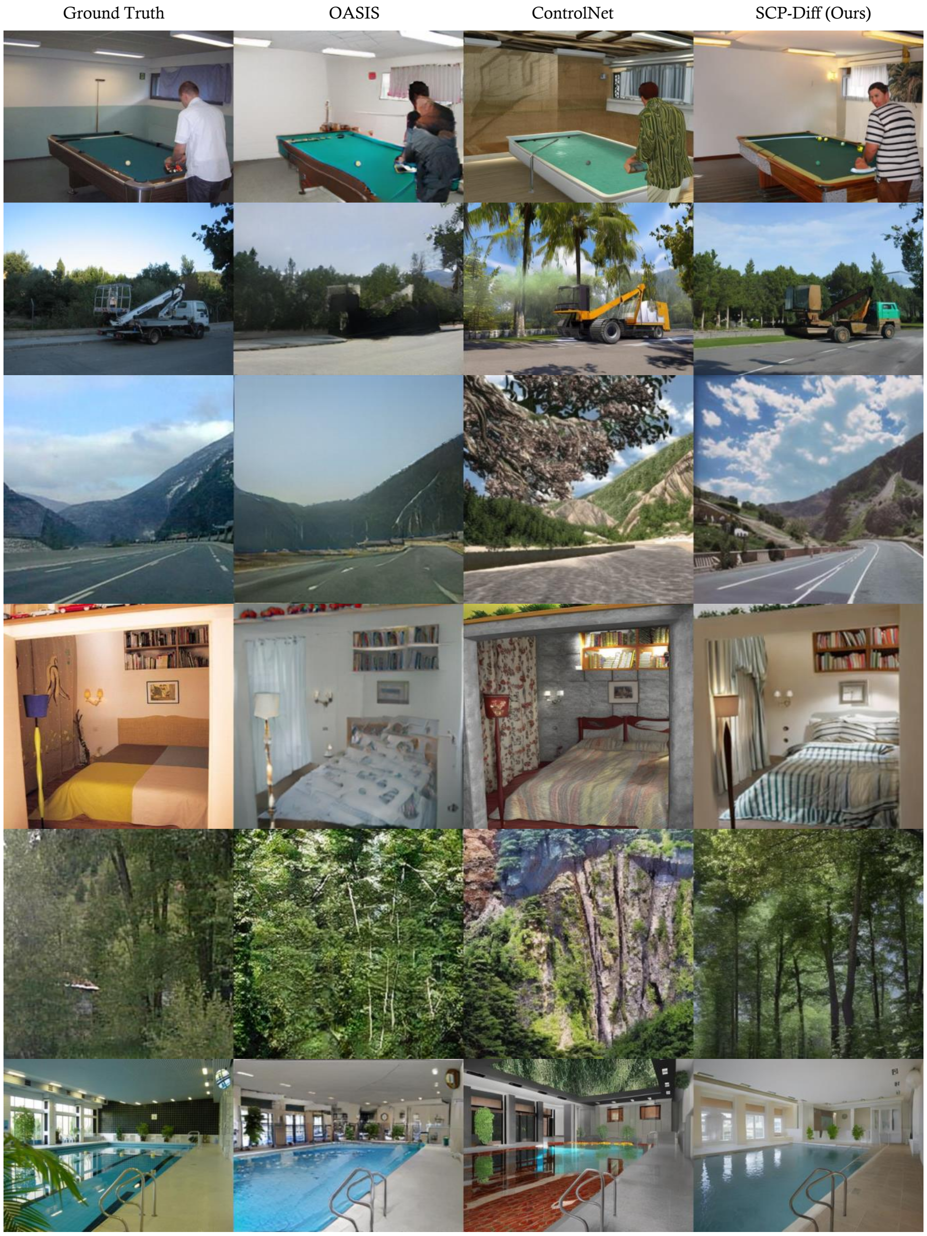

Fig. 4: Qualitative Results on ADE20K dataset. (cont.)

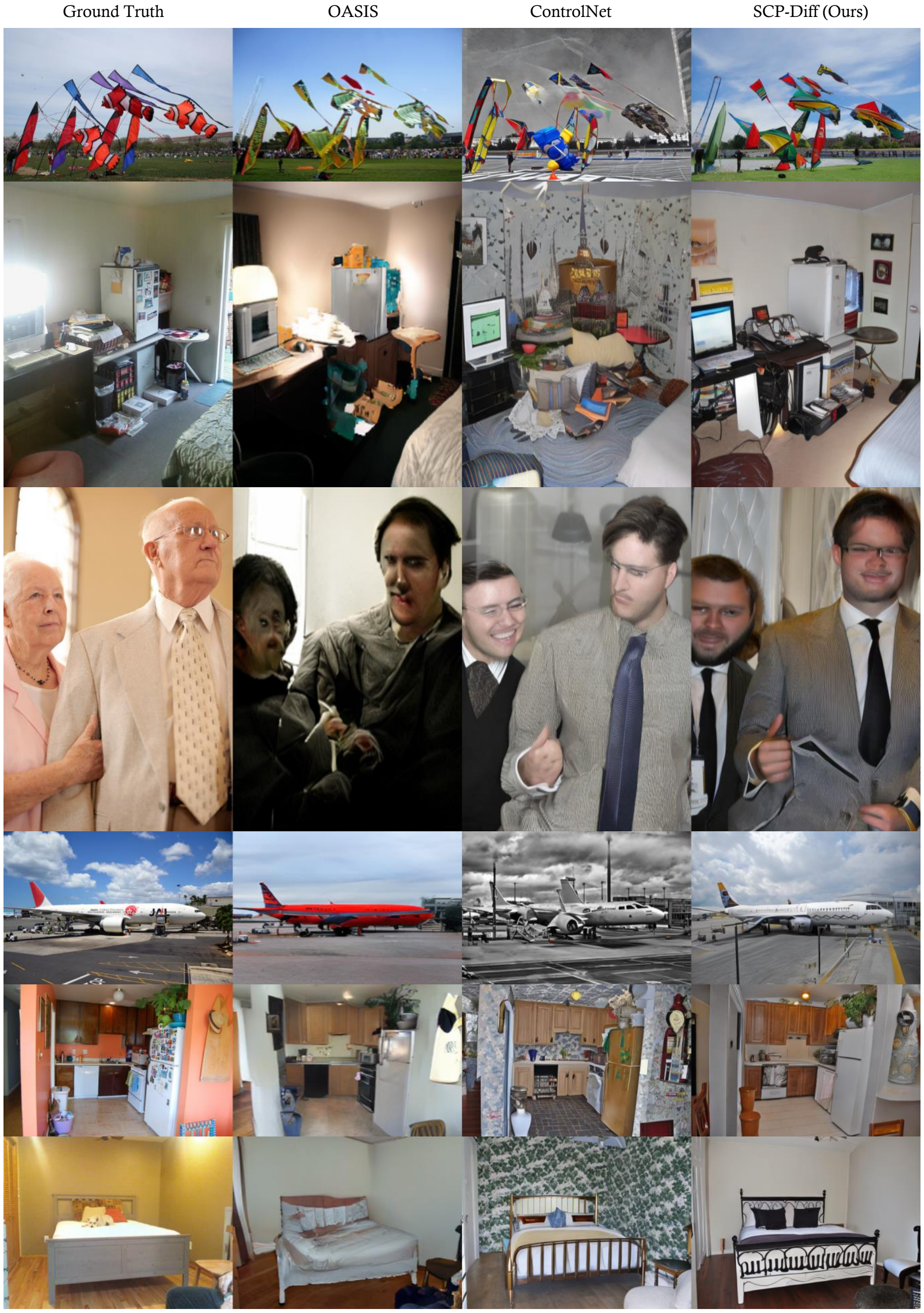

Fig. 5: Qualitative Results on COCO-Stuff dataset. (cont.)

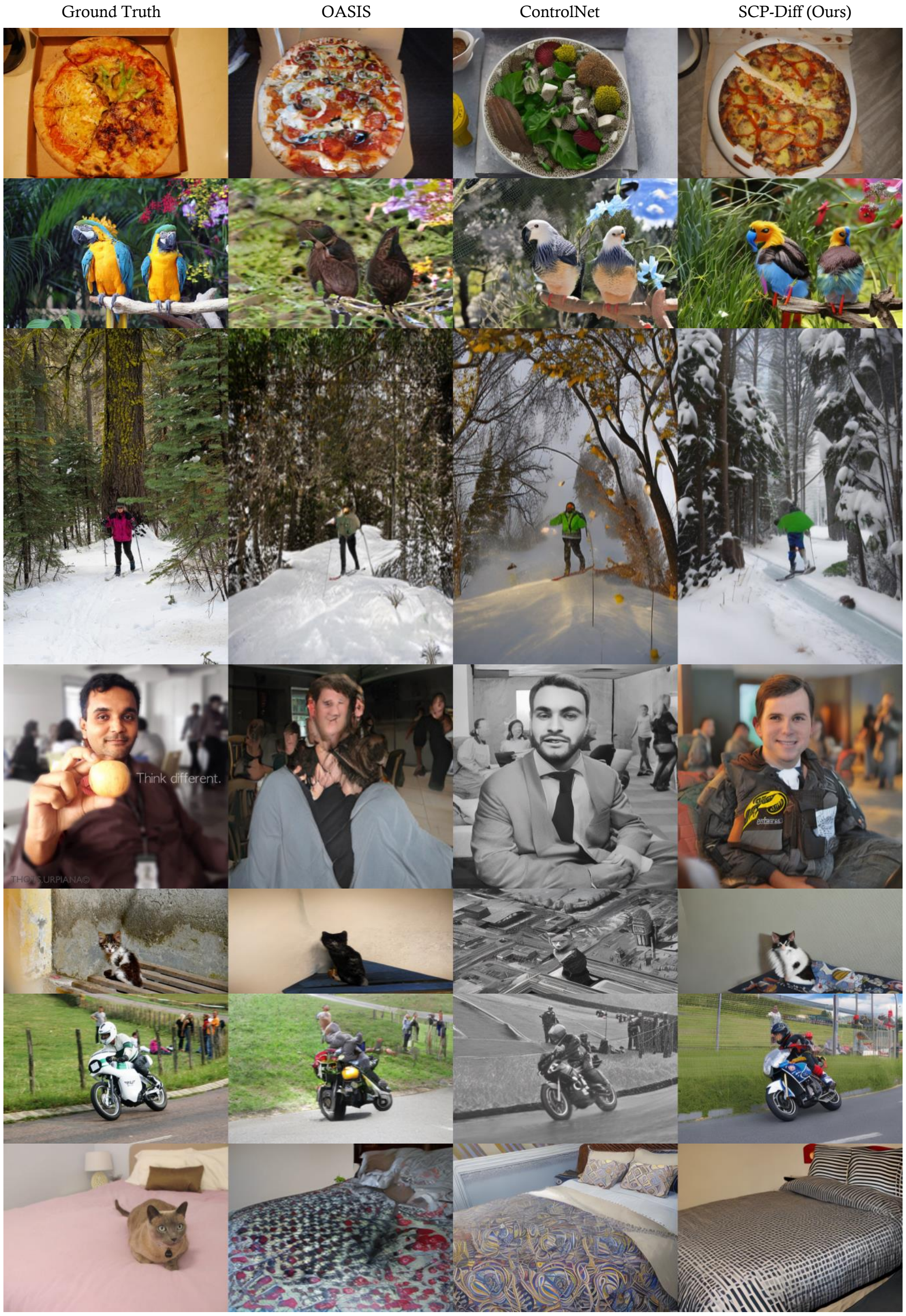

Fig. 6: Qualitative Results on COCO-Stuff dataset.