Modeling Image Composition for Complex Scene Generation Supplementary Material

Zuopeng Yang[1*] Daqing Liu[2*] Chaoyue Wang[3] Jie Yang[1†] Dacheng Tao[2,3]
[1]Shanghai JiaoTong University [2]JD Explore Academy, JD.com [3] The University of Sydney
{yzpeng, jieyang}@sjtu.edu.cn, chaoyue.wang@outlook.com, {liudq.ustc, dacheng.tao}@gmail.com

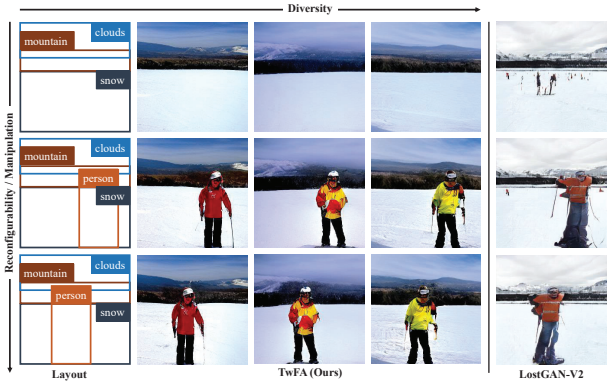

Figure 1. Examples of reconfigurable and diverse results (Section 6.1). Each row shows diverse generated images from the same layout on the left. Each column shows effects of the reconfiguration/manipulation by adding or moving objects. Compared with LostGANs-V2, TwFA synthesizes cleaner snow and better human structure.

6. Appendix

In this appendix, we first demonstrate the model's reconfigurable and diverse generation ability in Section 6.1. Next, we perform visual comparison with more SoTA methods in Section 6.2. Then, we provide more L2I Examples in Section 6.3 and more Few-shot L2I Examples in Section 6.4. Finally, the limitations and broader impacts will be discussed in Section 6.5

6.1. Reconfigurable and Diverse Generation

Reconfigurability. TwFA is reconfigurable and easy to manipulate as shown in each column of Figure 1. We can find that TwFA successfully maintains the styles after manipulating the layouts, *e.g.*, adding or moving objects. The reason is that focal attention insulates the patch-level interaction, *i.e.*, each patch only attends on related patches inner the same object and will not be affected by other unrelated patches.

Diversity. TwFA generates diverse images by multinomial resampling strategy, *i.e.*, we randomly sample each patch token according to its probability distribution $p(s_i)$ in Eq.(2). This multinomial resampling strategy introduces uncertainty into the generation process, leading to generation diversity. As shown in each row of Figure 1, TwFA can generate clothes of various colors and mountains of different styles.

6.2. Visual Comparison with More SoTAs

We perform visual comparison with more SoTA methods, including CNN-based (LostGAN-V2 [6], Context-L2I [2]) and Transformer-based (HCSS [4]), in Figure 3 and Figure 4. Compared with existing SoTA methods, the proposed TwFA can synthesize **1)** More reasonable object-level relationships, *e.g.*, person-surfboard, bird-bird, and person-motorcycle; **2)** Clearer patch-level instance structures, *e.g.*, oven, bird, and motorcycle; **3)** Refiner pixel-level textures, *e.g.*, rock, grass, and pavement.

6.3. More L2I Examples

More samples generated by TwFA are shown in Figure 5 – Figure 9. According to these results, several advantages of TwFA are observed: **1)** As illustrated in Figure 5, the reflection on the bus windows/river makes the images to look more realistic. **2)** The different textures of fur contribute to the fidelity for different animals, such as the cow, cat, and bird, shown in Figure 6. **3)** The well-generated structure for an object also reduces the unreality, (*i.e.*, bus, giraffe, and train). **4)** The shadow (*i.e.*, on the ground) conforms to the laws of physics, such as the last image in the sixth row of Figure 5. All of these advances make the generated images more realistic.

6.4. More Few-shot L2I Examples

To further demonstrate the advance of our method in few-shot complex scene generation, we conducted more experiments with different novel categories. The results are illustrated in Figure 2.

Settings. Few-shot complex scene generation consists of

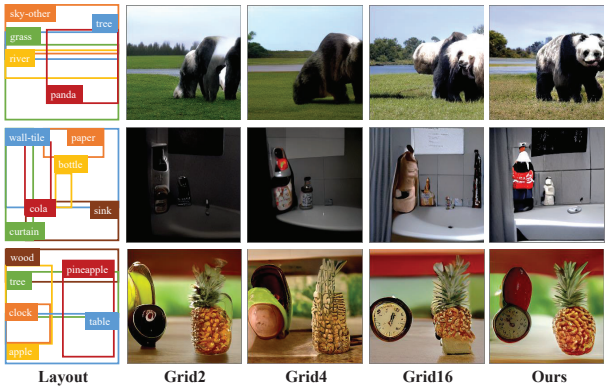

Figure 2. Examples of few-shot results (Section 6.4). The novel classes from the first row to the last one are the panda, cola, and pineapple, whose positions are annotated with red rectangles in layouts. TwFA outperforms all the baseline model with finer structures and details.

two steps: **1)** The first step is to prepare training data with annotated novel classes. To simplify the image annotation process, we randomly select several segmented novel objects and attach them into the images of COCO-stuff. Therefore, we can just annotate the bounding box positions of the novel classes. The rest annotations of the attached images derive from the dataset. **2)** The second step is to execute the few-shot implementation based on the models trained on the full COCO-stuff [1] dataset. To show the extreme situation, all the experiments are conducted in a setting of 2-shot. The novel classes include balloon, panda, penguin, Christmas tree, cola, and pineapple.

Comparisons. In this section, we discuss the few-shot result comparisons between our TwFA and other baseline models: **1)** As illustrated in Figure 2, our method can learn better instance structures, such as the balloon, panda, and pineapple. As for the penguin, the beak is generated more vividly than other baseline models. **2)** Our model not only learns how to synthesize the structure of an object, but also learns the interaction with the surroundings for the novel objects (*i.e.*, the penguin's shadow on the snow). **3)** Our model can reduce the interference of other stuff objects. In detail, Grid16 fails to synthesize the Christmas tree with the background of trees. Meanwhile, when generating the cola bottle, the disturbance from the surrounding patches leads to generating an arbitrary bottle instead of a cola bottle. In conclusion, TwFA outperforms all the baseline models with finer structures and details.

6.5. Limitations and Broader Impacts.

Limitations. Though the proposed TwFA achieves impressive performance on (few-shot) L2I tasks, it relies on the composition prior from given layouts which may not be available in other complex scene generation tasks, *e.g.*, textual descriptions [3] and scene graphs [5]. How to extend our method to those tasks remains open. We leave this as a promising direction to explore in the future.

Broader impacts. The proposed method synthesizes images based on learned patterns of the training dataset and as such will reflect biases in those data, including ones with negative societal impacts. The model may generate inexistent images with unexpected content. These issues warrant further research and consideration when generating images based on this work.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, 2018. 2

[2] Sen He, Wentong Liao, Michael Ying 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, 3, 4

[3] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Semantic object accuracy for generative text-to-image synthesis. *TPAMI*, 2020. 2

[4] Manuel Jahn, Robin Rombach, and Björn Ommer. High-resolution complex scene synthesis with transformers. In *CVPRW*, 2021. 1, 3, 4

[5] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *CVPR*, 2018. 2

[6] Wei Sun and Tianfu Wu. Learning layout and style reconfigurable gans for controllable image synthesis. *TPAMI*, 2021. 1, 3, 4

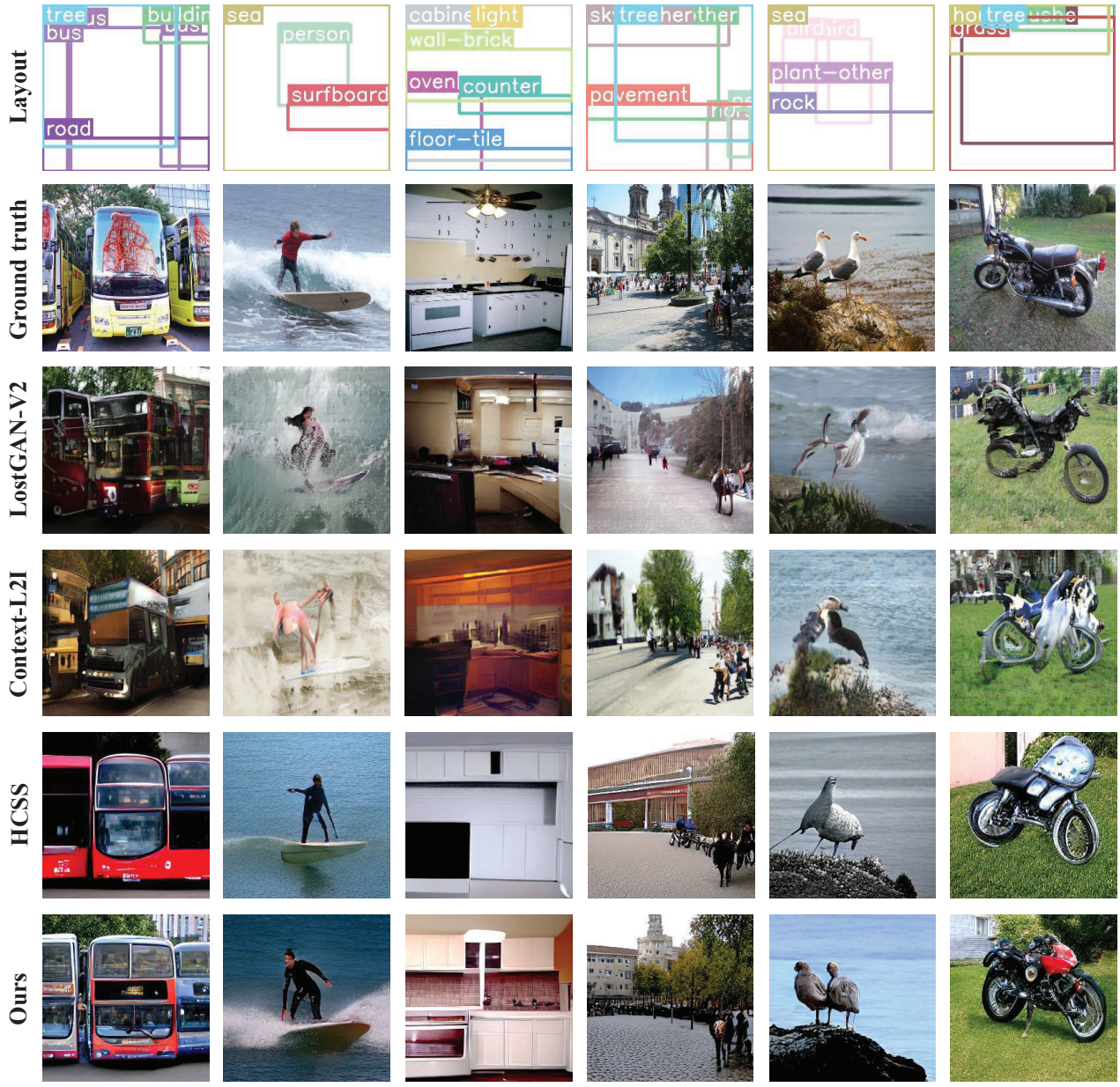

Figure 3. Comparisons with SoTAs (Section 6.2). Our method is compared against the most representative baseline model LostGAN-V2 [6], the existing state-of-the-art model Context-L2I [2], and the transformer-based method HCSS [4]. For all different scenes, TwFA outperforms the state-of-the-art model with more reasonable relationships, finer instance structures, and textures.

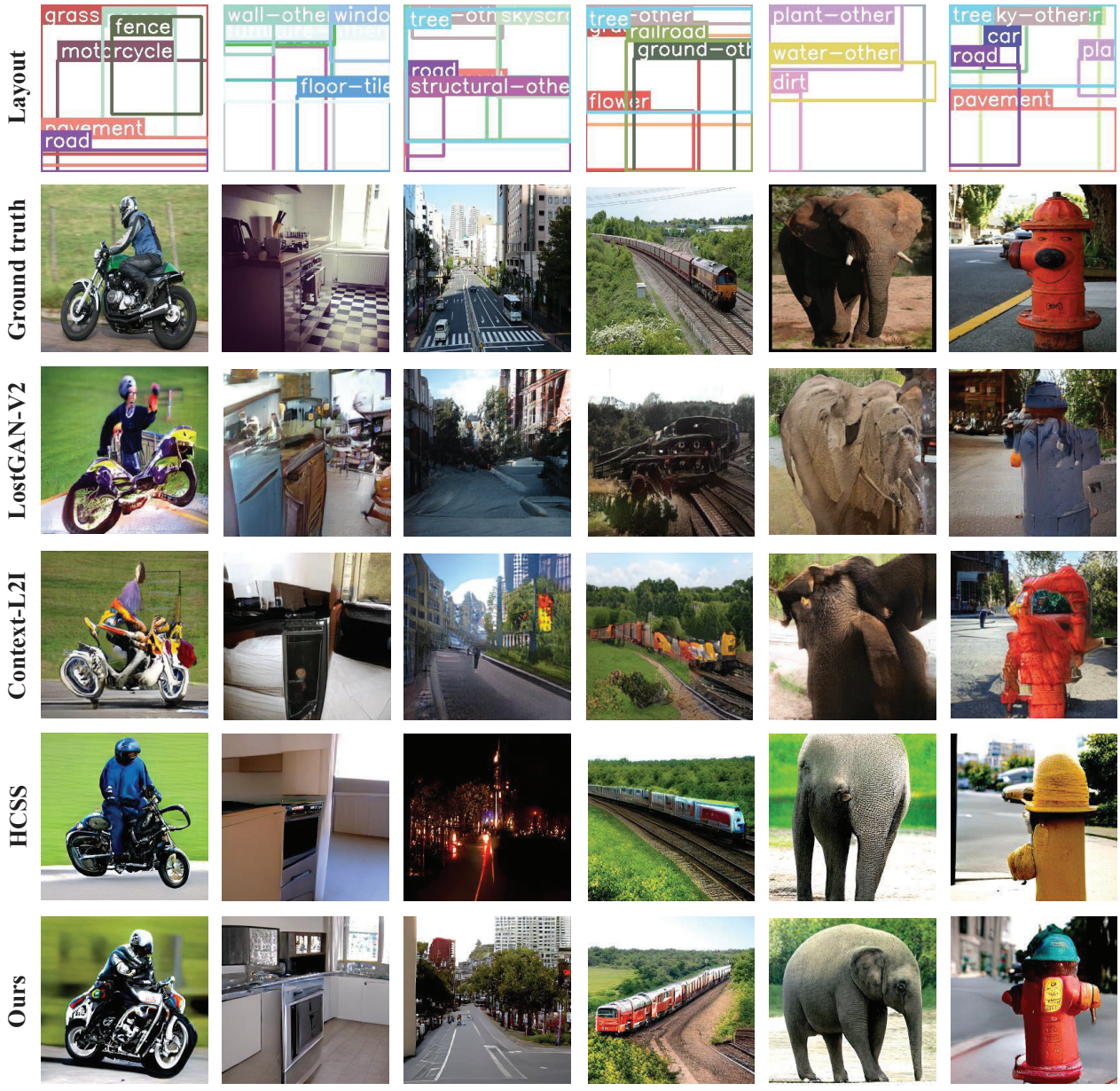

Figure 4. Comparisons with SoTAs (Section 6.2). Our method is compared against the most representative baseline model LostGAN-V2 [6], the existing state-of-the-art model Context-L2I [2], and the transformer-based method HCSS [4]. For all different scenes, TwFA outperforms the state-of-the-art model with more reasonable relationships, finer instance structures, and textures.

Figure 5. More L2I examples (Section 6.3). All images are generated by the proposed TwFA according to the given layout on the left.

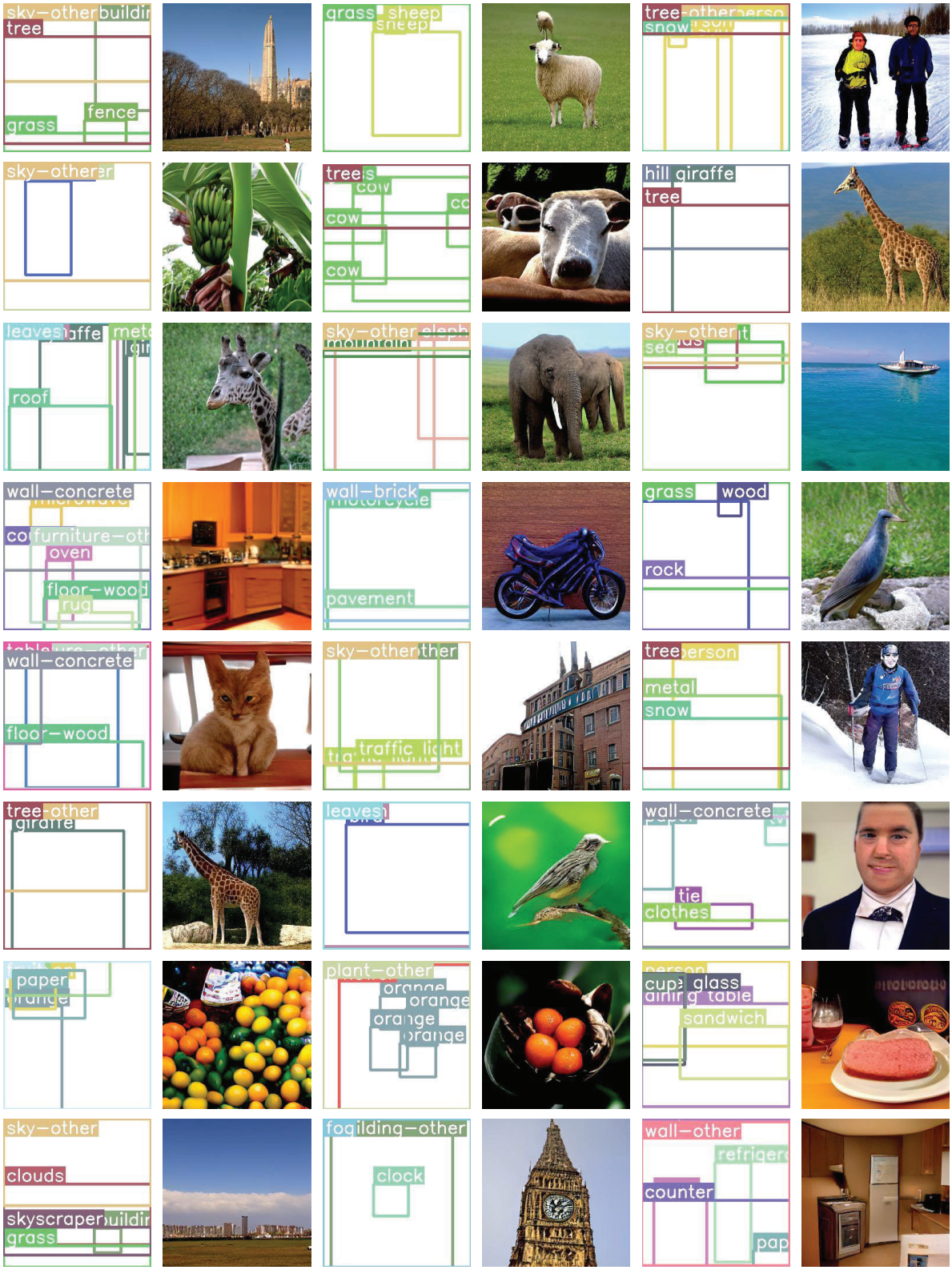

Figure 6. More L2I examples (Section 6.3). All images are generated by the proposed TwFA according to the given layout on the left.

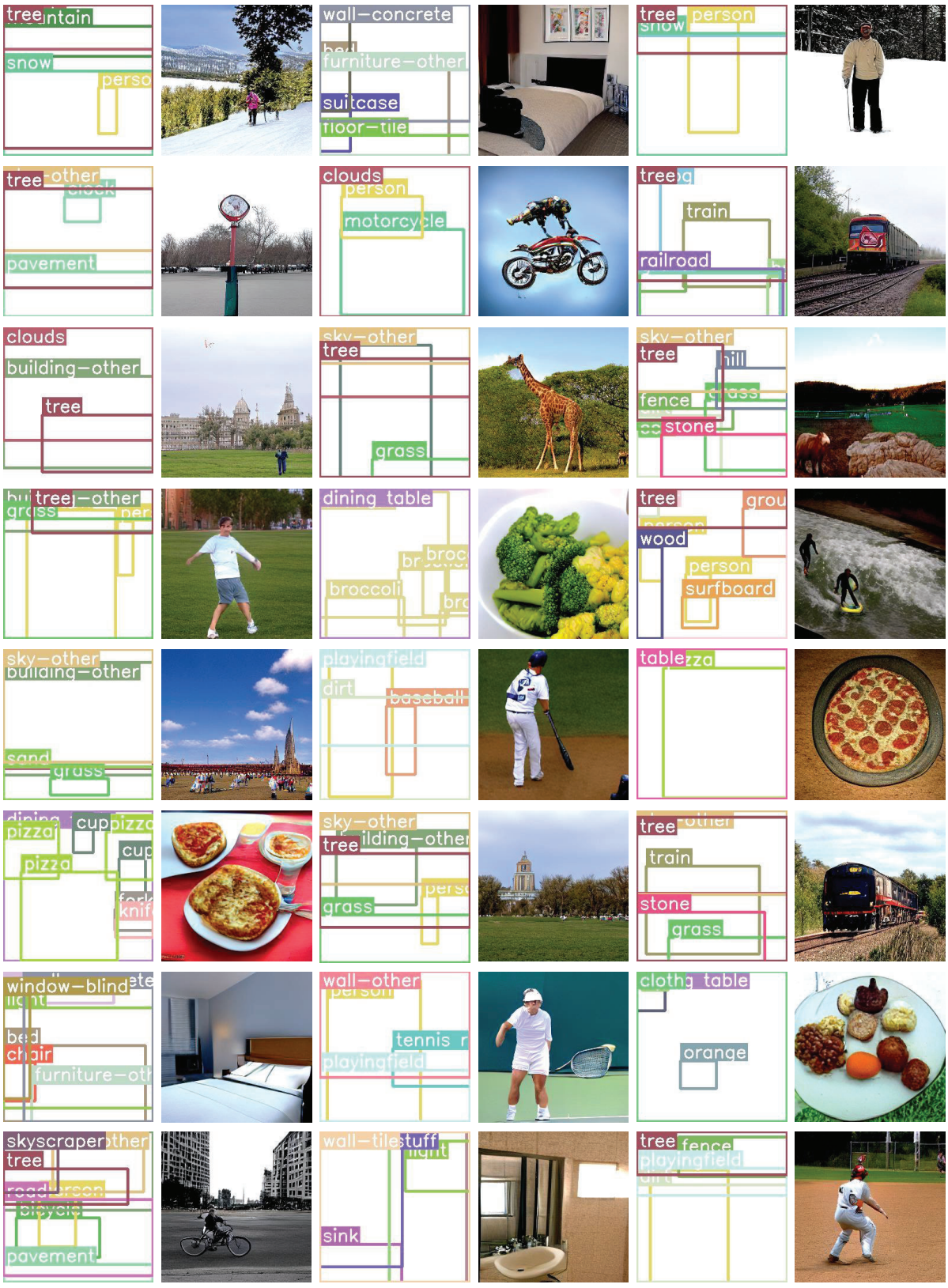

Figure 7. More L2I examples (Section 6.3). All images are generated by the proposed TwFA according to the given layout on the left.

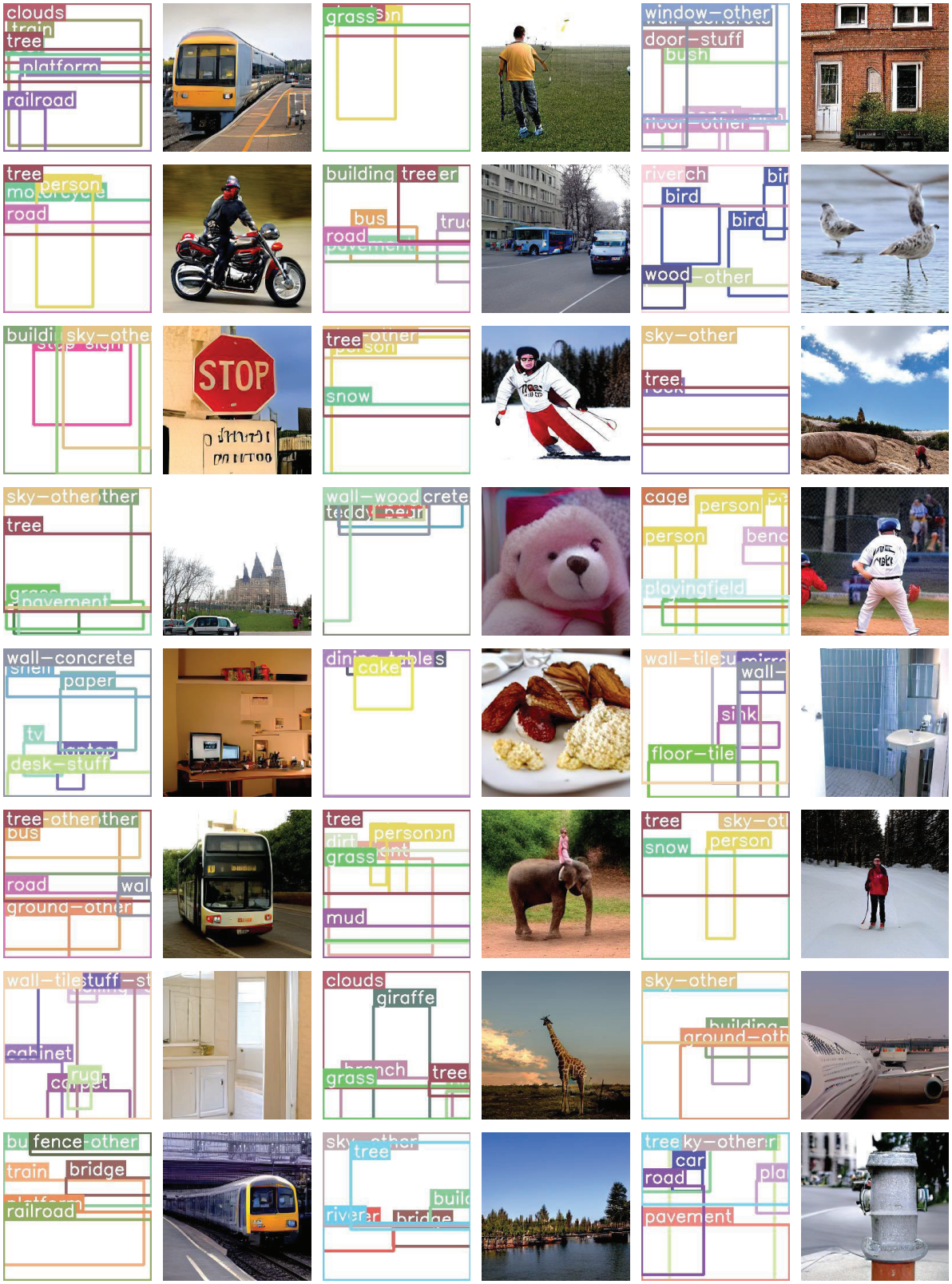

Figure 8. More L2I examples (Section 6.3). All images are generated by the proposed TwFA according to the given layout on the left.

Figure 9. More L2I examples (Section 6.3). All images are generated by the proposed TwFA according to the given layout on the left.