Multi-scale Contrastive Learning for Complex Scene Generation (Supplementary Material)

Hanbit Lee Youna Kim Sang-goo Lee
Seoul National University, Seoul, Korea
{skcheon,anna9812,sglee}@europa.snu.ac.kr

Type	Layer	Output Shape
-	Image	$256 \times 256 \times 3$
backbone	ResBlk	$256 \times 256 \times 128$
backbone	ResBlk	$128 \times 128 \times 256$
backbone	ResBlk	$64 \times 64 \times 512$
backbone	ResBlk	$32 \times 32 \times 512$
backbone	ResBlk	$16 \times 16 \times 512$
branch: shared	1×1 ResBlk	$16 \times 16 \times 512$
branch: disc	1×1 Conv	$16 \times 16 \times 1$
branch: proj	1×1 Conv	$16 \times 16 \times 256$
backbone	ResBlk	$8 \times 8 \times 512$
branch: shared	1×1 ResBlk	$8 \times 8 \times 512$
branch: disc	1×1 Conv	$8 \times 8 \times 1$
branch: proj	1×1 Conv	$8 \times 8 \times 256$
backbone	ResBlk	$4 \times 4 \times 512$
backbone	Flatten	8192
backbone	Linear	512
branch: disc	Linear	1
branch: proj	Linear	256

Table 1. Discriminator Architecture of MsConD.

1. Network Architecture

Table 1 shows the architectural details of MsConD discriminator. Our discriminator is built upon the backbone resnet-based discriminator used in StyleGAN2 [4]. We use branches to process the feature map at each level, where each branch consists of three components: a shared block, a discriminator head and a projection head. The shared block consists of 3 1×1 convolutional layers with residual connections. The shared block translates a feature map in the backbone network into an intermediate feature map of the same size. The intermediate feature map is then projected into two different outputs each by a discriminator head and a projection head, where both head layers are implemented with 1×1 convolutional layers. The discriminator head is a single convolutional layer, while the projection head consists of two convolutional layers, i.e., Conv-ReLU-Conv. We set the channel dimension of the projection output, i.e., C_p, as 256. We use ReLU activation for all layers in branches.

2. Additional Evaluation

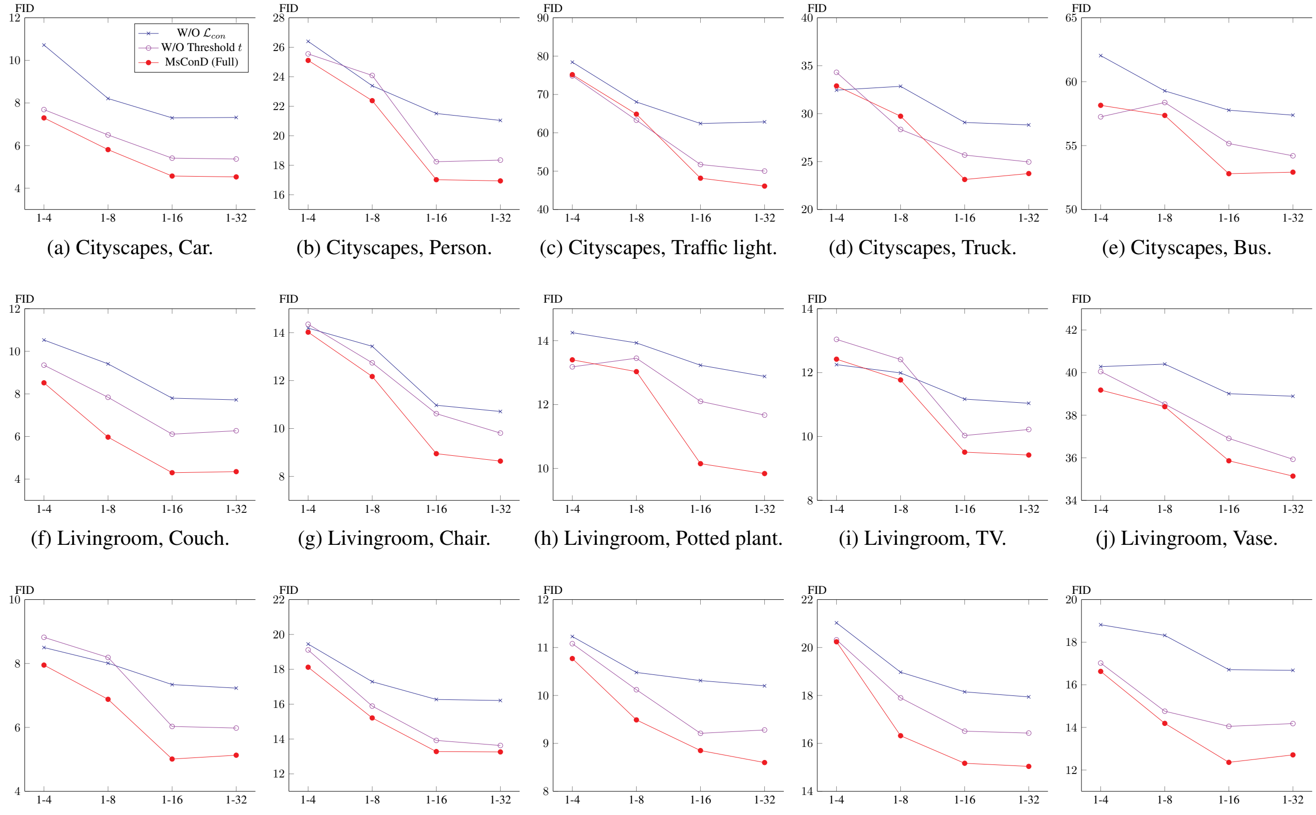
2.1. Additional Ablation Result

Figure 1 shows quantitative ablation result for each object category. We observe similar tendency as the scene-level ablation result (see Figure 4 in the main paper). The generation performance increases as more feature maps from the backbone layers are utilized even when multi-scale contrastive learning is not applied. However, the performance is significantly improved as the contrastive learning is leveraged as an auxiliary task. The improvement has been consistent across various object categories validating the efficacy of MsConD in synthesizing local objects.

2.2. Analysis on Training Dynamics.

To further understand the training behavior of MsConD, we investigate the statistics of discriminator logits for real and fake images during the training process. Figure 2 shows the result on Cityscapes. For StyleGAN2, the logit distributions overlap during the initial training period and then gradually move away from each other. Therefore, as the training progresses, the discriminator becomes overly confident and fails to provide meaningful feedback to the generator, resulting in degraded synthesis quality. To mitigate the overfitting of discriminator, previous studies mainly focus on developing differentiable image augmentations [2, 6]. Our findings indicate the problem could be substantially alleviated by utilizing multi-scale adversarial feedback.

Figure 2 (b-d) presents the results when local discriminator feedback is incorporated by our proposed discriminator. As shown, the logit distributions of real and fake samples remain within a close range for the entire training period, indicating that the discriminator can continuously provide informative feedback without overfitting. Meanwhile,



(a) Cityscapes, Car. (b) Cityscapes, Person. (c) Cityscapes, Traffic light. (d) Cityscapes, Truck. (e) Cityscapes, Bus.

(f) Livingroom, Couch. (g) Livingroom, Chair. (h) Livingroom, Potted plant. (i) Livingroom, TV. (j) Livingroom, Vase.

(k) Kitchen, Oven. (l) Kitchen, Chair. (m) Kitchen, Microwave. (n) Kitchen, Potted plant. (o) Kitchen, Refrigerator.

Figure 1. **Quantitative ablation result for each object category.** Comparison results in terms of object-level FID under different model configurations. For each configuration, we plot the results when different scales of feature maps are utilized. For example, '1-8' means that the model uses feature maps whose height is from 1 to 8.

we could observe that the fake logits for higher frequency part, i.e., $D^{16}_{disc}(x)$, tend to be unstable with large deviations. This instability stems from large structural variations of high frequency patterns in complex scenes. Figure 2 (e-g) shows that the auxiliary representation learning effectively stabilizes the feedback signal, in turn further improves the synthesis quality.

3. Additional Samples

Additional samples generated with MsConD. In Figure 3, we show additional samples generated by MsConD. Each row shows samples containing different objects of each object category. We mark the object bounding boxes detected by the object detector to emphasize the synthesis quality of individual objects.

Additional samples for comparison. For comparisons to the state of the art models, we provide more uncurated samples generated by different models in Figure 4, 5, 6. Compared to baselines, MsConD produces convincing results of more realistic scene images with improved local details.

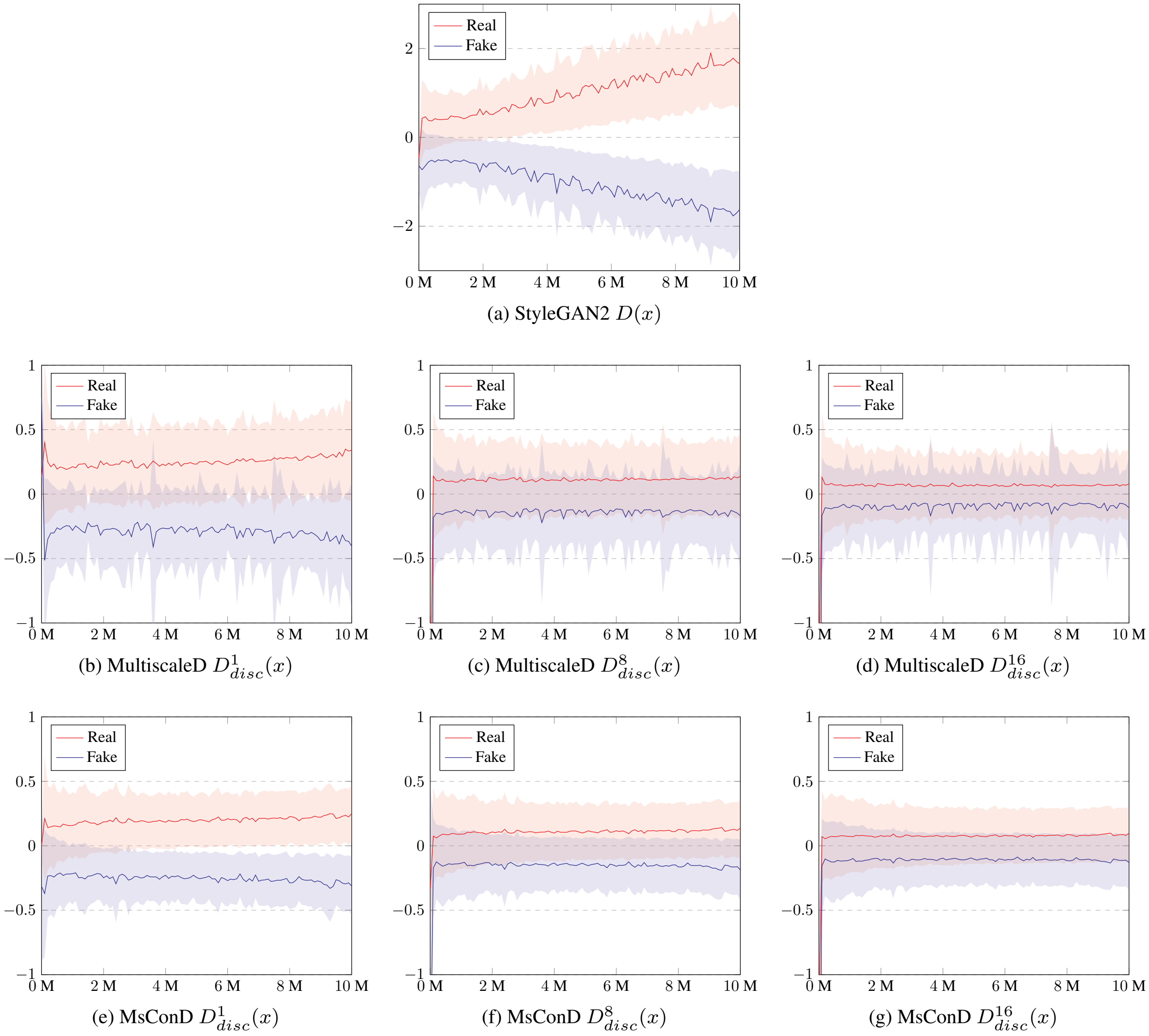
References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[2] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proc. NeurIPS*, 2020.

[3] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[4] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.



(a) StyleGAN2 $D(x)$

(b) MultiscaleD $D^1_{disc}(x)$

(c) MultiscaleD $D^8_{disc}(x)$

(d) MultiscaleD $D^{16}_{disc}(x)$

(e) MsConD $D^1_{disc}(x)$

(f) MsConD $D^8_{disc}(x)$

(g) MsConD $D^{16}_{disc}(x)$

Figure 2. **Training progress on Cityscapes.** Evolution of discriminator logits during training in (a) StyleGAN2, (b-d) MsConD without contrastive learning (MultiscaleD), and (e-g) MsConD. Here, the superscripts in each discriminator output notation represent the height of the output map at that scale, i.e., H_l. The horizontal axis of each plot corresponds to the training iterations (in number of images). The bold line indicates mean value of logits and the transparent area indicates the deviation.

[5] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[6] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *Advances in Neural Information Processing Systems*, 33:7559–7570, 2020.

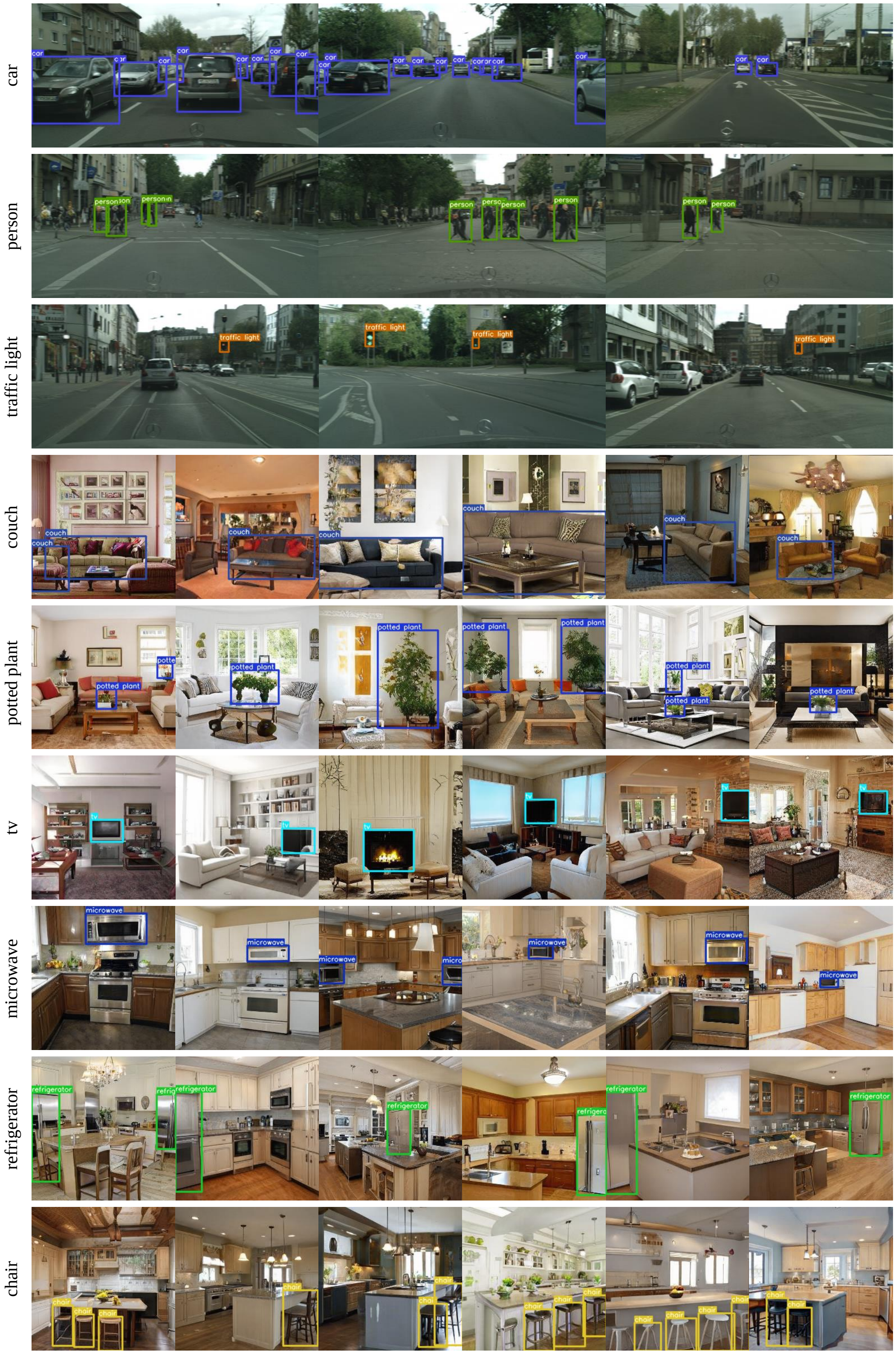
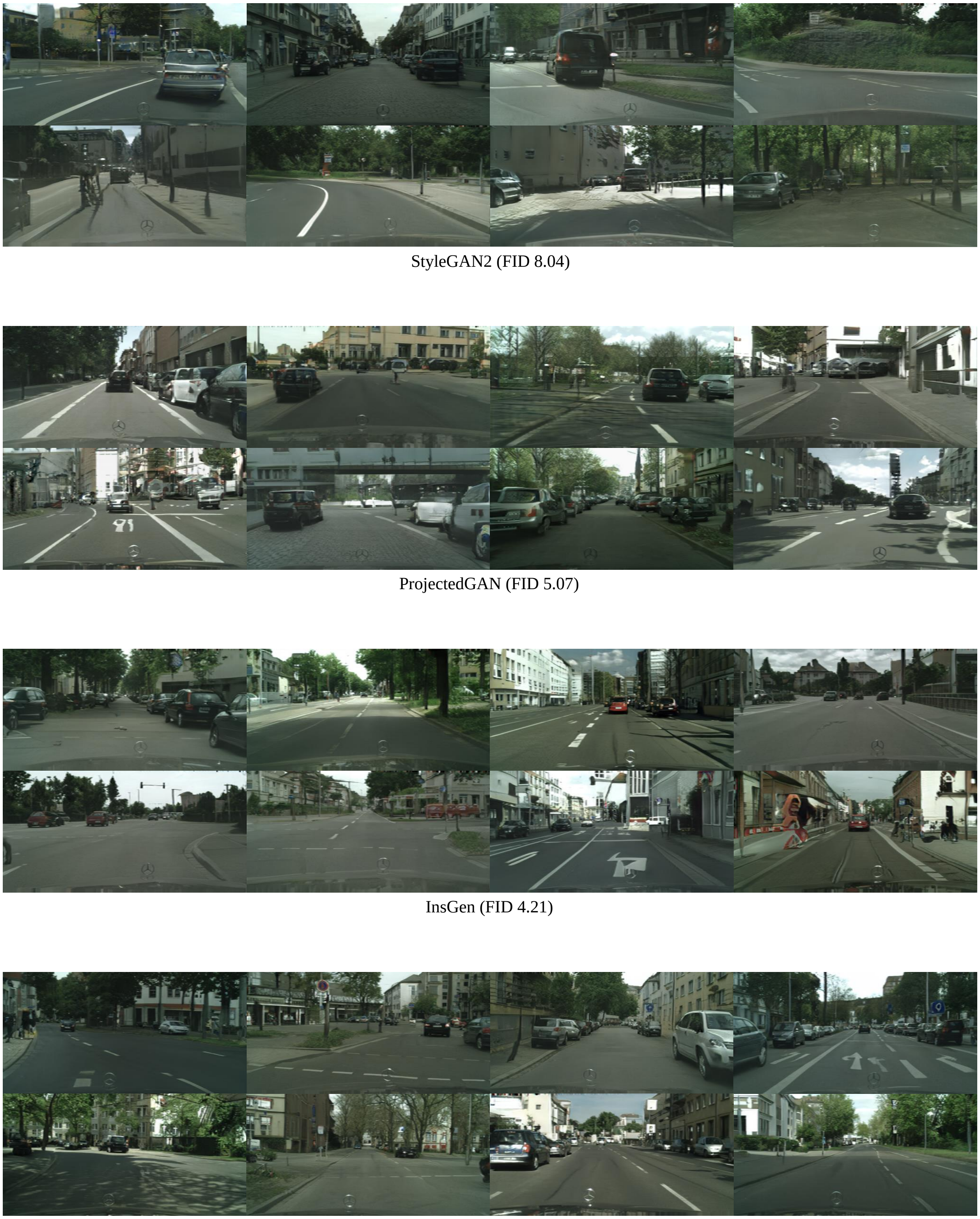


Figure 3. **Samples Generated with MsConD.** Each row shows samples containing objects of each object category. All images are generated with truncation trick following [3, 2]. We recommend zooming in to inspect the synthesis quality of individual objects.



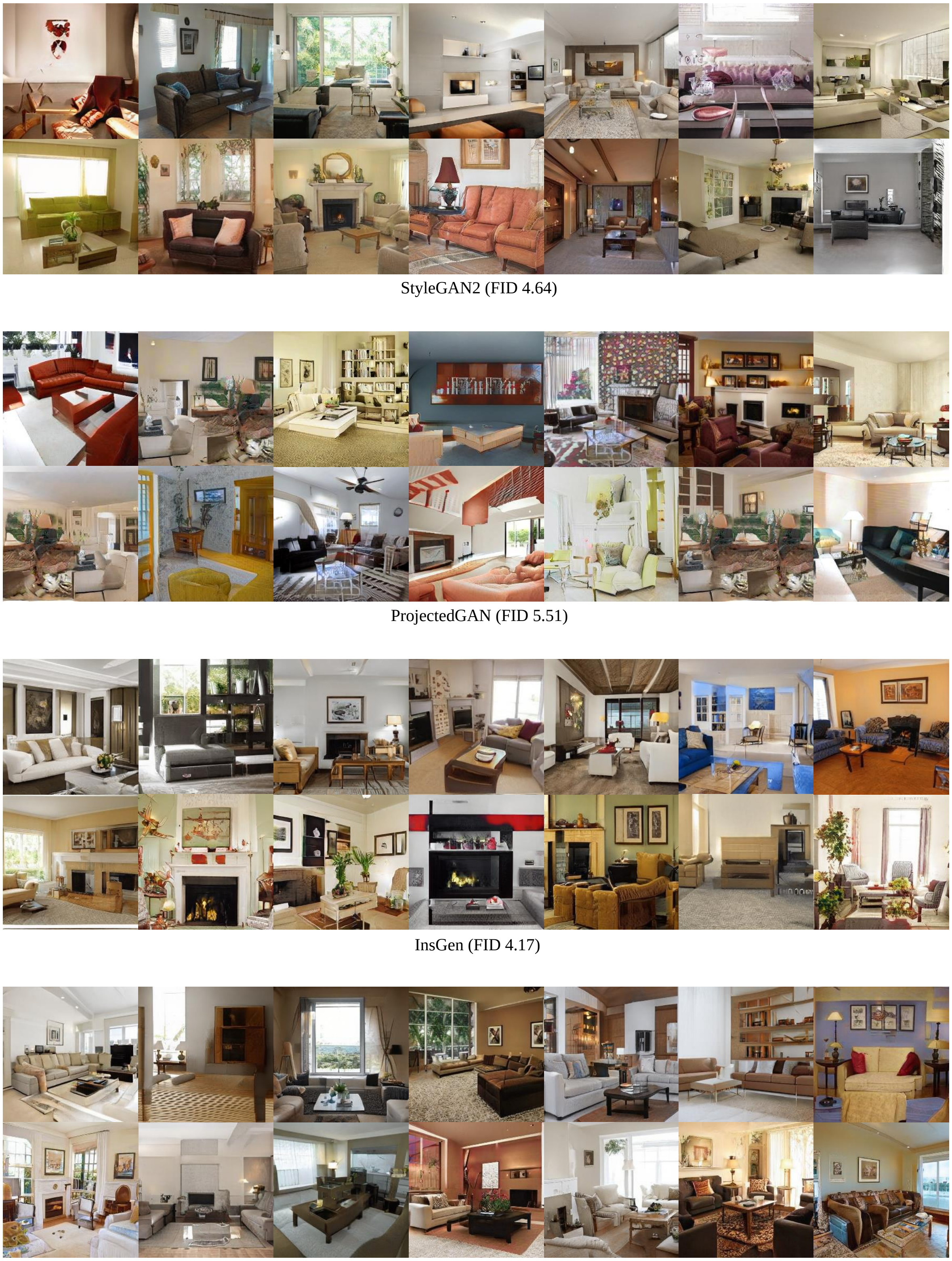
StyleGAN2 (FID 8.04)

ProjectedGAN (FID 5.07)

InsGen (FID 4.21)

MsConD (Ours) (FID 2.63)

Figure 4. **Uncurated Samples for Cityscapes [1].** All images are randomly sampled with truncation trick following [3, 2]. We recommend zooming in to compare scene details.



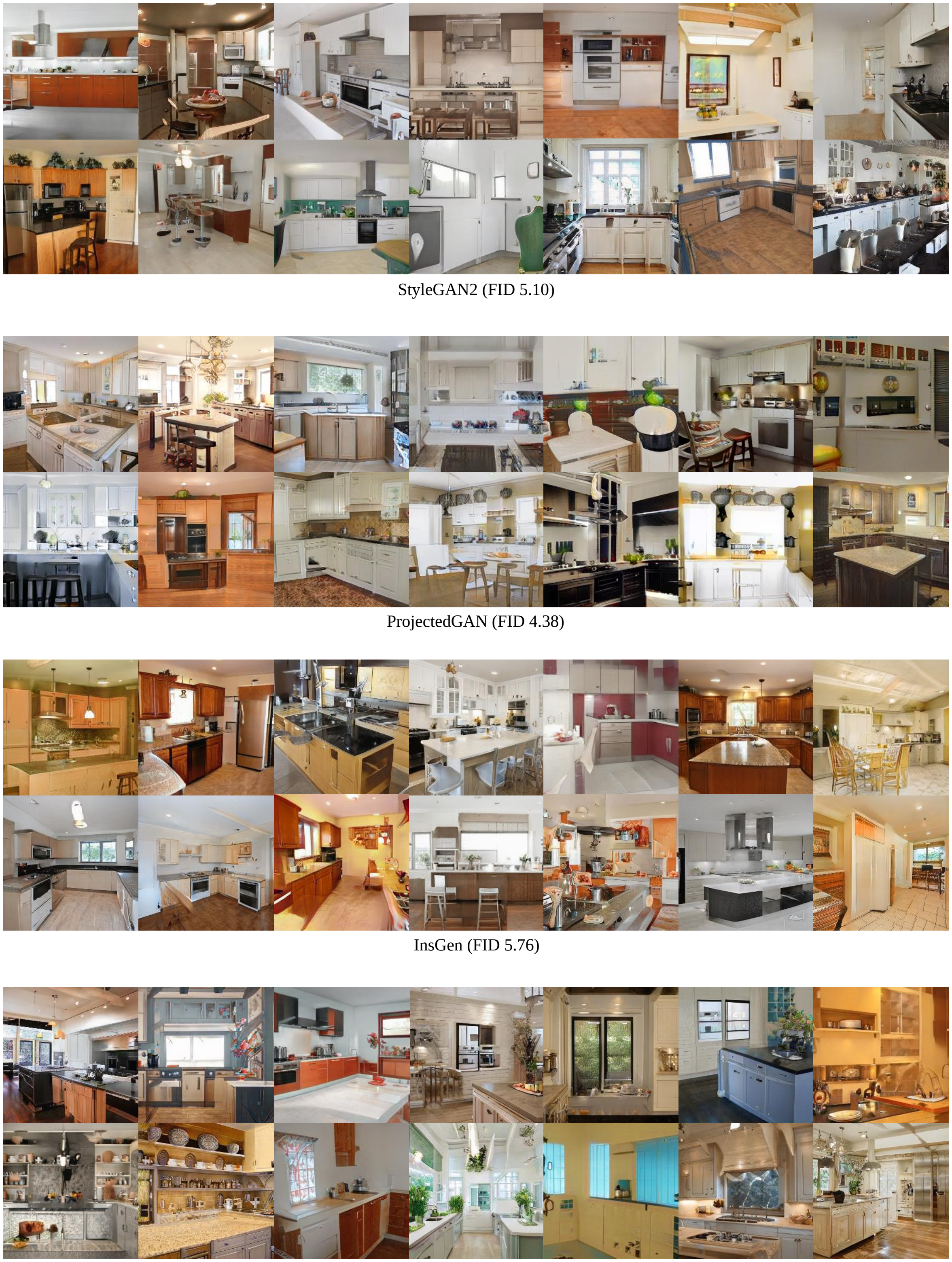
StyleGAN2 (FID 4.64)

ProjectedGAN (FID 5.51)

InsGen (FID 4.17)

MsConD (Ours) (FID 2.73)

Figure 5. **Uncurated Samples for Livingroom [5].** All images are randomly sampled with truncation trick following [3, 2]. We recommend zooming in to compare scene details.



StyleGAN2 (FID 5.10)

ProjectedGAN (FID 4.38)

InsGen (FID 5.76)

MsConD (Ours) (FID 2.88)

Figure 6. **Uncurated Samples for Kitchen [5].** All images are randomly sampled with truncation trick following [3, 2]. We recommend zooming in to compare scene details.