GIRAFFE HD: A High-Resolution 3D-aware Generative Model
Supplementary Material

Yang Xue[1] Yuheng Li[2] Krishna Kumar Singh[3] Yong Jae Lee[2]
[1]UC Davis [2]UW–Madison [3]Adobe Research

1. Full Loss Expression

For a given generator-discriminator pair $\{G, D\}$, the overall objective function can be formalized as

$$\begin{aligned} &L(G, D) = \\ &\mathbb{E}_{z_a^k, z_s^k \sim \mathcal{N}, \xi \sim p_\xi, T^k \sim p_T}[f(D(G(\{z_a^k, z_s^k, T^k\}_k, \xi)))] \\ &+ \mathbb{E}_{I \sim p_D}[f(-D(I)) - \frac{\lambda}{2}\|\nabla D(I)\|^2] \\ &+ \beta_1 L_{bbox} + \beta_2 L_{cvg} + \beta_3 L_{bin} \end{aligned} \tag{1}$$

where $f(t) = -\log(1+\exp(-t))$, $\lambda = 10$, p_D indicates the data distribution, and $\beta_1, \beta_2, \beta_3$ are dataset specific. L_{bbox}, L_{cvg}, and L_{bin} are as defined in the main paper.

2. Mutual Background Similarity (MBS) Details

We denote the generator to evaluate as G, which takes as input randomly sampled foreground parameters $P_{fg} \sim p_{fg}$ and background parameters $P_{bg} \sim p_{bg}$ to generate an image I. We denote a pretrained semantic segmentation model DeepLabV3 ResNet101 [1] as R which takes an image I and outputs the semantic prediction map for I, which can then be converted into the background mask M. We compute the mutual background similarity (MBS) by first randomly sampling an image $I_1 = G(P_{fg_1}, P_{bg})$, then generating another image by sampling another $P_{fg_2} \sim p_{fg}$ while keeping P_{bg} fixed, $I_2 = G(P_{fg_2}, P_{bg})$. Then we compute the background masks for the two images $M_1 = R(I_1), M_2 = R(I_2)$ and the mask for the two images' mutual background area can be computed as $M_{multbg} = M_1 \cdot M_2$. We define that a pixel's RGB value has changed if one or more channels of the pixel's RGB value has changed over some small threshold η. Then the total number of pixels inside the mutual background area whose RGB value has changed is computed as

$$N = \sum\nolimits_{i \in M_{multbg}} \delta \tag{2}$$

where $\delta = \begin{cases} 0, & \text{if } \eta > |I_1[i][c] - I_2[i][c]|, c \in \{R, G, B\} \\ 1, & \text{otherwise} \end{cases}$

The image is normalized to [0,1] before feeding into R, and η is set to be $\frac{1}{255}$. Then the MBS for image pair $\{I_1, I_2\}$ is

$$MBS = \frac{N}{|M_{multbg}|} \times 100 \tag{3}$$

In Figures 1 and 2, we show the segmentations produced by DeepLabV3 ResNet101 [1] and the mutual background difference map for both GIRAFFE and GIRAFFE HD (ours) on FFHQ [4] and CompCar [6] datasets. For GIRAFFE HD on FFHQ, the mutual background difference mainly comes from the imprecision of the segmentation (as DeepLabV3 cannot properly segment thin, floating hair). For GIRAFFE HD on CompCar, the mutual background difference mainly comes from the segmentor not including the car's shadow as part of the foreground.

3. Dataset Details

Dataset parameters. We report the dataset-dependent camera elevation angle and valid object transformation parameters used for all the datasets in Table 1. We use the same dataset parameters as GIRAFFE for CompCar, FFHQ, LSUN Church and CelebA-HQ datasets (except for CompCar's vertical translation). Since GIRAFFE was not evaluated on AFHQ Cat, we use the same dataset parameters GIRAFFE uses for Cats [8].

4. Additional Qualitative Results

In Figs. 3 to 23, we show additional qualitative results on controllable scene generation on four datasets: CompCar [6], FFHQ [4], AFHQ Cat [2], LSUN Church [7]. Since the results on CelebA-HQ [3] are very similar to those on FFHQ, we do not show the CelebA-HQ results here. We also include GIRAFFE samples on the four datasets to enable direct comparison with our method. We show the highest resolution models that we've trained for each dataset:

	Number of Images	Object Rotation Range	Background Rotation Range	Camera Elevation Range	Horizontal Translation	Depth Translation	Vertical Translation	Object Scale	Field of View
CompCar [6]	136,726	360°	0°	10°	-0.12 - 0.12	-0.22 - 0.22	-0.06 - 0.08	0.8 - 1	10°
FFHQ [4]	70,000	70°	0°	10°	-	-	-	-	10°
AFHQ Cat [2]	5,558	70°	0°	10°	-	-	-	-	10°
LSUN Church [7]	126,227	360°	0°	0°	-0.15 - 0.15	-0.15 - 0.15	-	0.8 - 1	30°
CelebA-HQ [3]	30,000	90°	90°	10°	-	-	-	-	10°

Table 1. **Dataset parameters.** We report relevant parameters for all datasets.

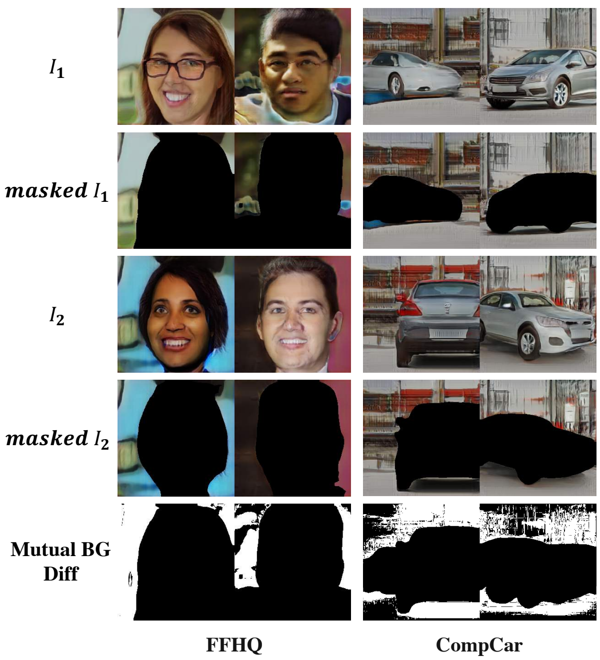


Figure 1. **GIRAFFE MBS Calculation.** DeepLabV3 background segmentations and mutual background differences (white pixels) used for computing MBS on GIRAFFE samples.

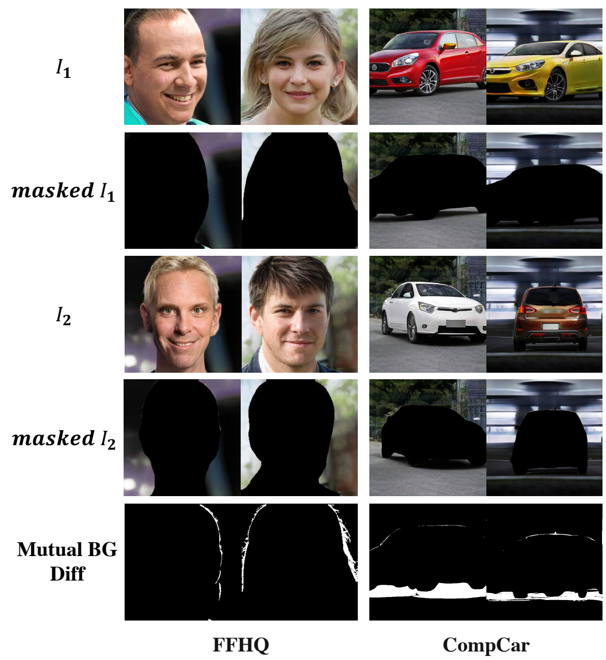


Figure 2. **GIRAFFE HD (ours) MBS Calculation.** DeepLabV3 background segmentations and mutual background differences (white pixels) used for computing MBS on our GIRAFFE HD samples.

CompCar at 512^2, FFHQ at 1024^2, AFHQ Cat at 256^2, and LSUN Church at 256^2.

References

[1] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. 2017. 1

[2] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 1, 2, 4, 6, 8, 10, 14, 18

[3] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 1, 2

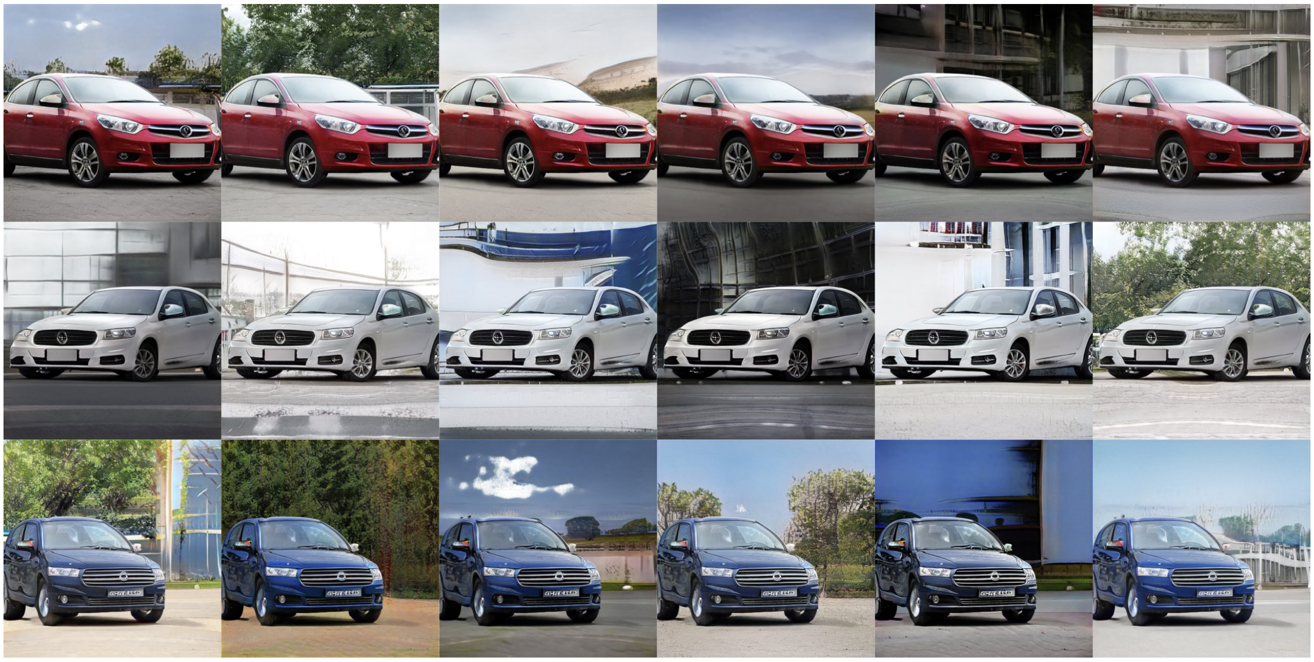
[4] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 2, 3, 5, 7, 9, 13, 17

[5] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021. 16, 17, 18, 19

[6] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *CVPR*, 2015. 1, 2, 3, 5, 7, 9, 11, 12, 16

[7] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv*, abs/1506.03365, 2015. 1, 2, 4, 6, 8, 10, 11, 15, 19

[8] Weiwei Zhang, Jian Sun, and Xiaoou Tang. Cat head detection - how to effectively exploit shape and texture features. In *ECCV*, 2008. 1



(a) CompCar 512^2 change background

(b) FFHQ 1024^2 change background

Figure 3. **Controllable Image Synthesis.** Changing background results on CompCar [6] and FFHQ [4]. Notice how the appearance of the foreground adapts to the changing background.

(a) AFHQ Cat 256^2 change background

(b) LSUN Church 256^2 change background

Figure 4. **Controllable Image Synthesis.** Changing background results on AFHQ Cat [2] and LSUN Church [7]. Notice how the appearance of the foreground adapts to the changing background. We also observe that for datasets where the foreground object does not have great variation in appearance (e.g., LSUN Church), the refine foreground renderer tends to take more control over the final foreground object's appearance than the initial foreground renderer. In such cases, making changes to the background tends to change the foreground appearance more.

(a) CompCar 512^2 change appearance

(b) FFHQ 1024^2 change appearance

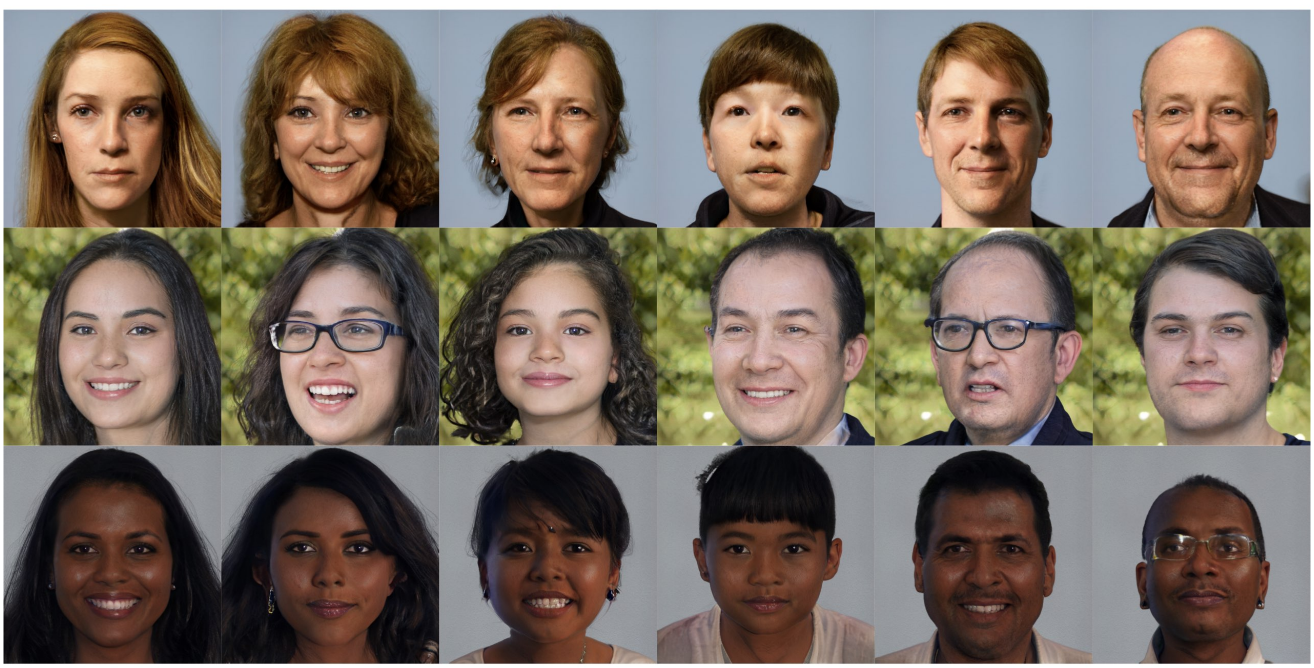
Figure 5. **Controllable Image Synthesis.** Changing appearance results on CompCar [6] and FFHQ [4].

(a) AFHQ Cat 256^2 change appearance

(b) LSUN Church 256^2 change appearance

Figure 6. **Controllable Image Synthesis.** Changing appearance results on AFHQ Cat [2] and LSUN Church [7]. As mentioned previously, for datasets where the foreground object does not have great variation in appearance (e.g., LSUN Church), the refine foreground renderer tends to take more control over the final foreground object's appearance than the initial foreground renderer. In such cases, making changes to the foreground appearance code tends to have relatively less effect on the appearance of the foreground object.

(a) CompCar 512^2 change shape



(b) FFHQ 1024^2 change shape

Figure 7. **Controllable Image Synthesis.** Changing shape results on CompCar [6] and FFHQ [4].

(a) AFHQ Cat 256^2 change shape

(b) LSUN Church 256^2 change shape

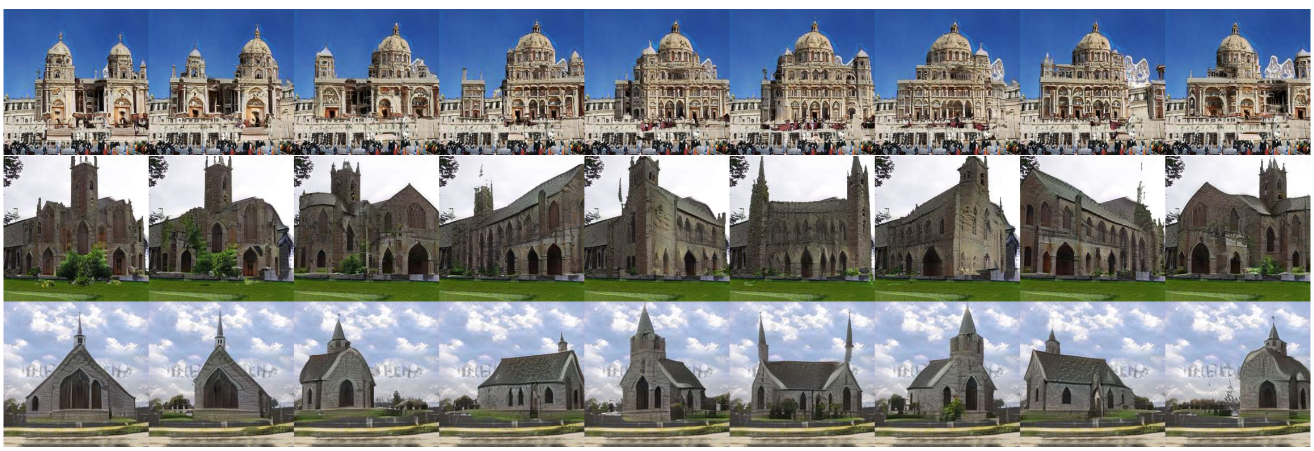
Figure 8. **Controllable Image Synthesis.** Changing shape results on AFHQ Cat [2] and LSUN Church [7].

(a) CompCar 512^2 rotation and camera elevation

(b) FFHQ 1024^2 rotation and camera elevation

Figure 9. **Controllable Image Synthesis.** Changing rotation and camera elevation results on CompCar [6] and FFHQ [4].

(a) AFHQ Cat 256^2 rotation



(b) LSUN Church 256^2 rotation

Figure 10. **Controllable Image Synthesis.** Changing rotation and camera elevation results on AFHQ Cat [2] and changing rotation results on LSUN Church [7] (the model is trained with a fixed camera elevation on the LSUN Church dataset). We observe that changing the camera elevation has little effect on the AFHQ Cat results. We attribute this to its small dataset size.

(a) Depth Translation

(b) Horizontal Translation

(c) Vertical Translation

(d) Scaling

Figure 11. **Controllable Image Synthesis.** Translation and scaling results on CompCar [6] and LSUN Church [7].

Figure 12. **Comprehensive Outputs.** Intermediate and final output images for CompCar [6] 512^2.

Figure 13. **Comprehensive Outputs.** Intermediate and final output images for FFHQ [4] 1024^2.

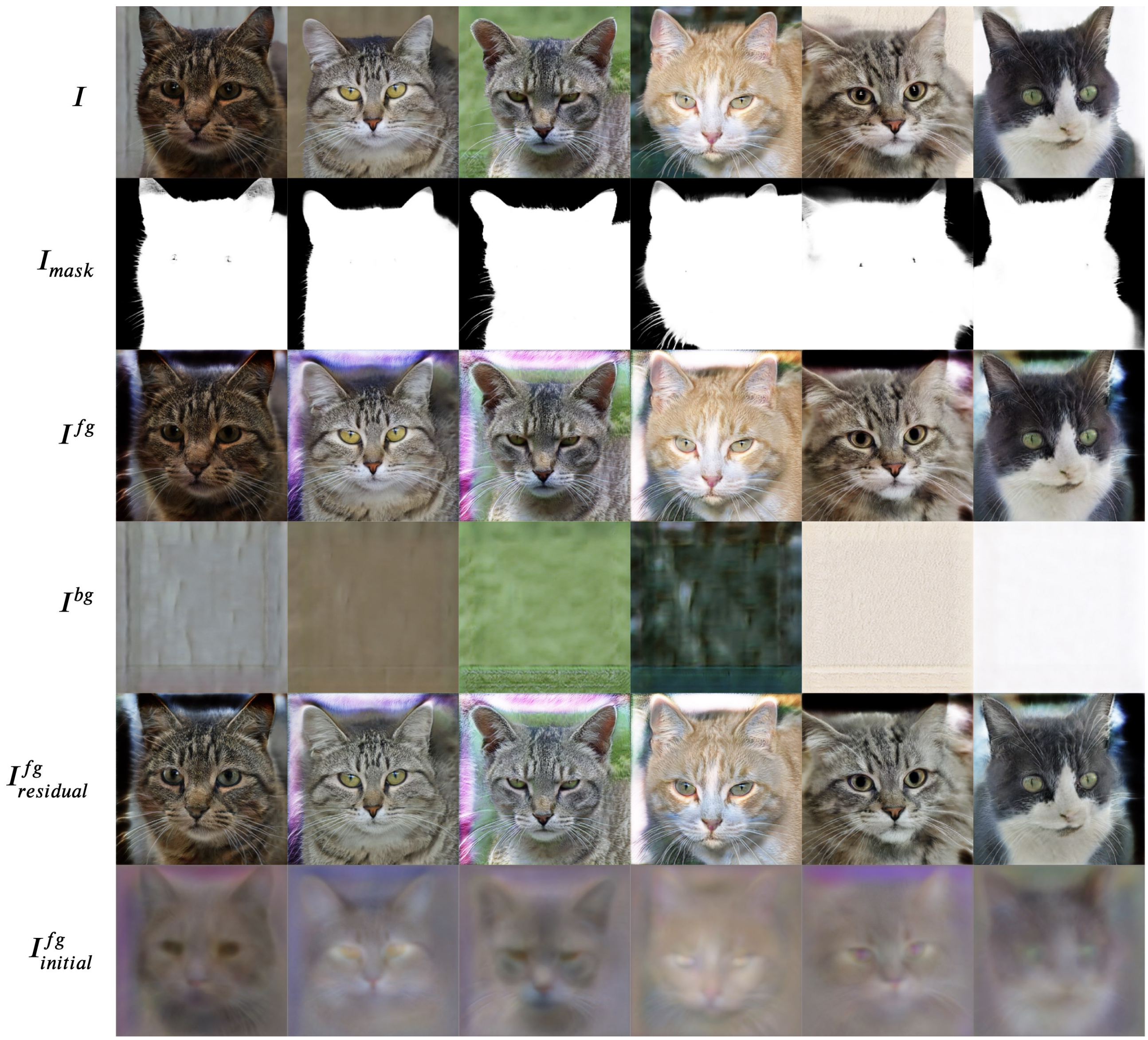


Figure 14. **Comprehensive Outputs.** Intermediate and final output images for AFHQ Cat [2] 256^2.

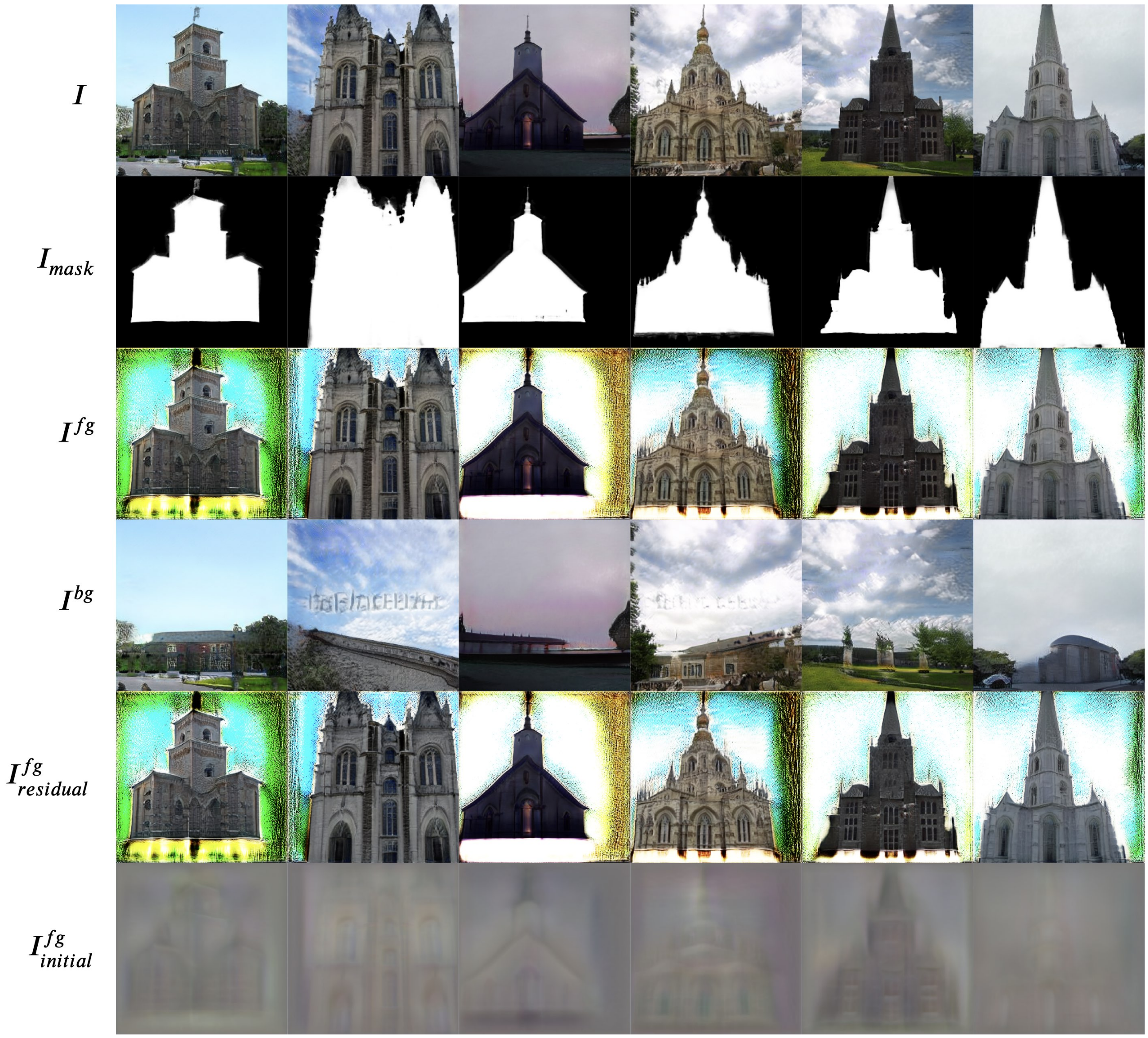


Figure 15. **Comprehensive Outputs.** Intermediate and final output images for LSUN Church [7] 256^2.

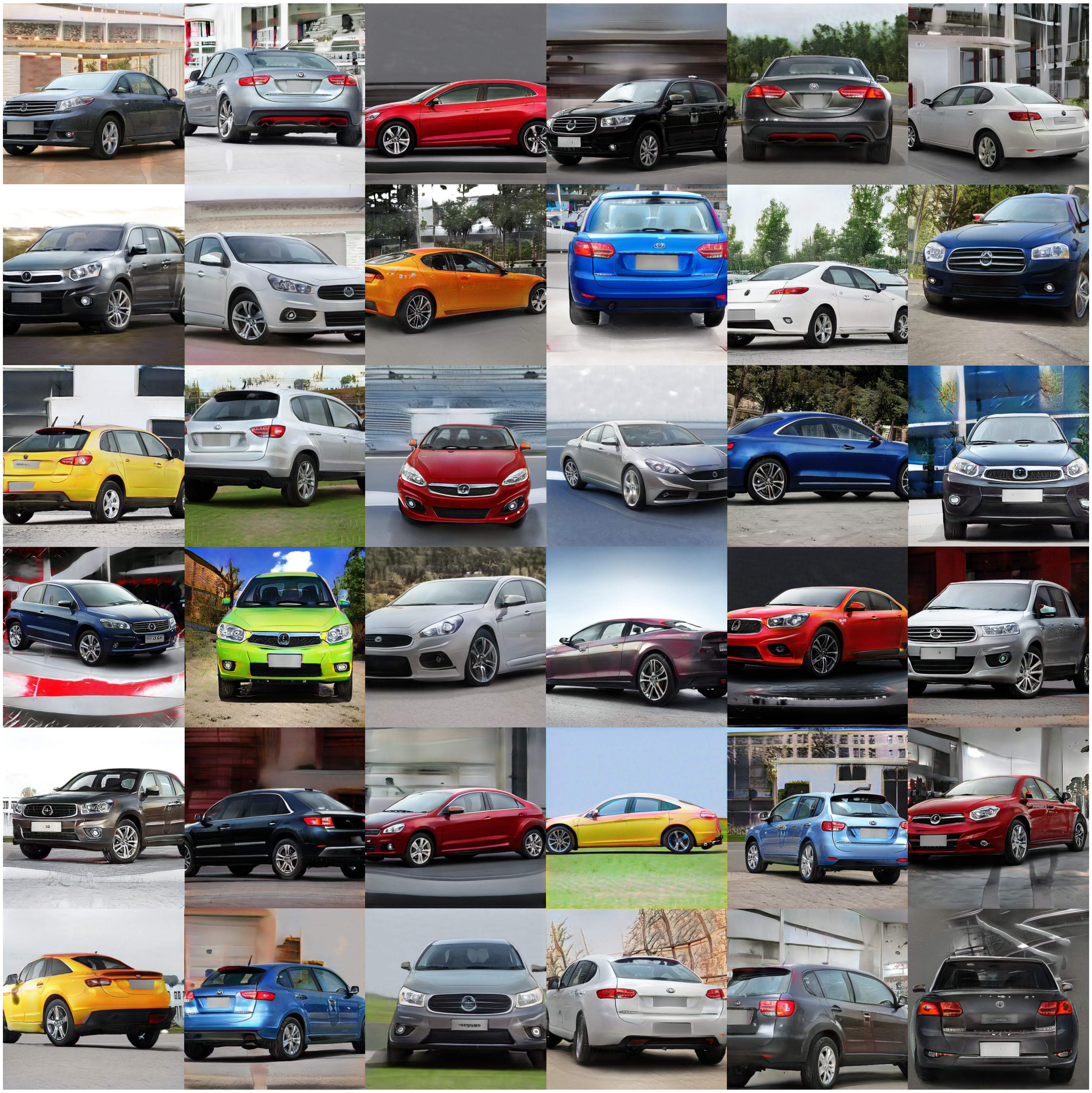


Figure 16. **Our samples.** GIRAFFE HD samples on CompCar [6] 512^2 .

Figure 17. **GIRAFFE [5] samples.** GIRAFFE samples on CompCar 256^2.

Figure 18. **Our samples.** GIRAFFE HD samples on FFHQ [4] 1024^2.

Figure 19. **GIRAFFE [5] samples.** GIRAFFE samples on FFHQ 256^2.

Figure 20. **Our samples.** GIRAFFE HD samples on AFHQ Cat [2] 256^2.

Figure 21. **GIRAFFE [5] samples.** GIRAFFE samples on AFHQ Cat 256^2.

Figure 22. **Our samples.** GIRAFFE HD samples on LSUN Church [7] 256^2.

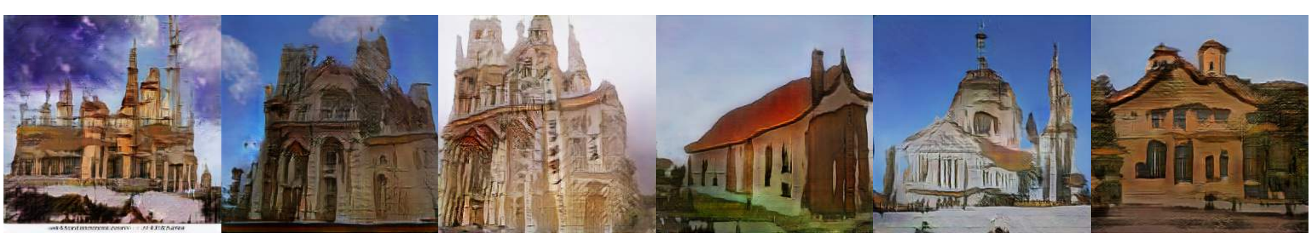


Figure 23. **GIRAFFE [5] samples.** GIRAFFE samples on LSUN Church 256^2.