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

Supplementary Material

Yang Xue¹  Yuheng Li²  Krishna Kumar Singh³  Yong Jae Lee²  
¹UC Davis  ²UW–Madison  ³Adobe Research

1. Full Loss Expression

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

\[
L(G, D) = \mathbb{E}_{z_h^k, z_q^k \sim N, \xi \sim p_D, T \sim p_T} \left[ f(D(G(z_h^k, z_q^k, T, \xi))) \right] \\
+ \mathbb{E}_{I \sim p_D} \left[ f(-D(I)) - \frac{\lambda}{2} \| \nabla D(I) \|_2^2 \right] \\
+ \beta_1 L_{bbox} + \beta_2 L_{cvg} + \beta_3 L_{bin}
\]  

(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 \( I_1 = G(P_{fg1}, P_{bg}) \), then generating another image by sampling another \( P_{fg2} \sim p_{fg} \) while keeping \( P_{bg} \) fixed, \( I_2 = G(P_{fg2}, P_{bg}) \). 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 \( \eta \). Then the total number of pixels inside the mutual background area whose RGB value has changed is computed as

\[
N = \sum_{i \in M_{multbg}} \delta
\]

(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 \( \eta \) is set to be \( \frac{1}{255} \). Then the MBS for image pair \( \{I_1, I_2\} \) is

\[
MBS = \frac{N}{|M_{multbg}|} \times 100
\]

(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:
### Table 1. Dataset parameters. We report relevant parameters for all datasets.

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

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

CompCar at 512², FFHQ at 1024², AFHQ Cat at 256², and LSUN Church at 256².

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.

---

### References


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.
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.
Figure 5. **Controllable Image Synthesis.** Changing appearance results on CompCar [6] and FFHQ [4].
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.
Figure 7. **Controllable Image Synthesis.** Changing shape results on CompCar [6] and FFHQ [4].
Figure 8. **Controllable Image Synthesis.** Changing shape results on AFHQ Cat [2] and LSUN Church [7].
Figure 9. **Controllable Image Synthesis.** Changing rotation and camera elevation results on CompCar [6] and FFHQ [4].
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.
Figure 11. **Controllable Image Synthesis.** Translation and scaling results on CompCar [6] and LSUN Church [7].
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] 1024².

Figure 19. **GIRAFFE [5] samples**. GIRAFFE samples on FFHQ 256².
Figure 20. Our samples. GIRAFFE HD samples on AFHQ Cat [2] 256².

Figure 21. GIRAFFE [5] samples. GIRAFFE samples on AFHQ Cat 256².
Figure 22. **Our samples.** GIRAFFE HD samples on LSUN Church [7] 256².

Figure 23. **GIRAFFE [5] samples.** GIRAFFE samples on LSUN Church 256².