# Repaint123: Fast and High-quality One Image to 3D Generation with Progressive Controllable Repainting

Junwu Zhang[1⋆] Zhenyu Tang[1⋆] Yatian Pang[1,3] Xinhua Cheng[1]
Peng Jin[1] Yida Wei[4] Xing Zhou[5] Munan Ning[1,2] Li Yuan[1,2⋆⋆]
[1]Peking University [2]Pengcheng Laboratory [3]National University of Singapore
[4]Wuhan University [5]RabbitPre

## 1 Evaluation on Multi-view Dataset

We choose the Google Scanned Object (GSO) dataset [1] and use 10 objects for a 120-view evaluation of the generated 3D objects with 3D ground truth. As shown in Table 1, the results indicate that our method is capable of generating high-quality 3D content with view consistency compared to the baselines. Though Table 1 shows similar PSNR and SSIM performance between Repaint123 and DreamGaussian, it primarily stems from shared geometry generated in the coarse stage. As our texture refinement strategy is performed on UV space in the fine stage, these pixel-level aligned metrics may not effectively demonstrate the superiority of our method when there is a geometric misalignment between our result in the coarse stage and ground-truth, due to the diversity of image-to-3D generation. Therefore, we also adopt two metrics for non-aligned similarity evaluation, CLIP Similarity and Contextual Distance, which confirm that Repaint123 achieves superior alignment with ground-truth textures.

| **Method\ Metric** | PSNR↑ | SSIM↑ | LPIPS↓ | CLIP↑ | Contextual↓ |
|---|---|---|---|---|---|
| Syncdreamer | 13.201 | 0.784 | 0.322 | 0.612 | 1.686 |
| Magic123 | 14.985 | 0.803 | 0.244 | 0.767 | 1.376 |
| Zero-123-XL | 15.118 | 0.813 | 0.229 | 0.761 | 1.334 |
| DreamGaussian | 15.391 | 0.814 | 0.237 | 0.736 | 1.407 |
| **Repaint123** | **15.393** | **0.814** | **0.214** | **0.812** | **1.319** |

**Table 1:** Multi-view quantitative comparison with state-of-the-art image-to-3D generation baselines on GSO dataset.

⋆ Equal contribution.
⋆⋆ Corresponding author.

| Dataset | Methods \ Metrics | CLIP-Similarity↑ | Context-Dis↓ | PSNR↑ | LPIPS↓ | Refine Time↓ |
|---|---|---|---|---|---|---|
| **RealFusion15** | Magic123 | 0.82 | 1.64 | 19.68 | 0.107 | 30min |
| | **Ours*** | **0.85** | **1.57** | **20.27** | **0.096** | **5min** |
| **Test-alpha** | Magic123 | 0.84 | 1.57 | 24.69 | 0.046 | 30min |
| | **Ours*** | **0.88** | **1.46** | **24.61** | **0.036** | **5min** |

**Table 2:** We show the comparison results with Magic123 in terms of CLIP-Similarity↑ / Contextual-Distance↓ / PSNR↑ / LPIPS↓ when we adopt NeRF representation for the coarse stage. **Bold** reflects the best performance. * indicates our NeRF-based Repaint123 replaces the refinement strategy in the fine stage of Magic123 with the proposed repainting approach and optimizes for the same amount of steps with MSE loss.


**Fig. 1:** Visual comparison between NeRF-based Repaint123 and Magic123.

## 2 Evaluation of NeRF-based Repaint123

As our repainting approach is plug-and-play for the refinement stage, we can change the representation in the coarse stage from Gaussian Splatting to NeRF. As presented in Table 2 and Figure 1, the generated 3D objects can be significantly improved by using our repainting method. Compared with Magic123 for the optimization time of the fine stage, our NeRF-based approach reaches a significant acceleration of **6** times, due to efficient texture refinement.

## 3 More Ablation Study

**Analysis of Prompts.** We conduct ablations on various prompts, including image prompt, text prompt, textual inversion, and empty prompt. As shown in Table 3, prompts significantly enhance both view consistency and generation quality compared to results obtained without prompts. The efficacy stems from the classifier-free guidance technique. Among various prompts, image prompts demonstrate superior performance, showcasing the superior accuracy of visual prompts over text prompts, including time-consuming optimization-based textual inverted prompts.

| **Prompt\ Metric** | PSNR↑ | LPIPS↓ | CLIP↑ | Contextual↓ |
|---|---|---|---|---|
| None | 19.02 | 0.102 | 0.79 | 1.60 |
| Text | 19.00 | 0.102 | 0.83 | 1.58 |
| Textual Inversion | 19.01 | 0.101 | 0.84 | 1.57 |
| **Image** | 19.00 | **0.101** | **0.85** | **1.55** |

**Table 3:** Ablation study on RealFusion15 dataset under various prompt conditions. Image prompt achieves superior performance.

**Analysis of Angular Interval.** Our examination, summarized in Table 4, identifies that a 40-degree angular interval achieves the best reference-view PSNR and LPIPS while a 60-degree angular interval achieves the best CLIP similarity. However, Figure 2 shows that 60-degree angular interval decreases the size of shared visible areas between neighboring views, raising the likelihood of multi-head problems during optimization. Consequently, a 40-degree angular interval was selected as optimal for the training process.

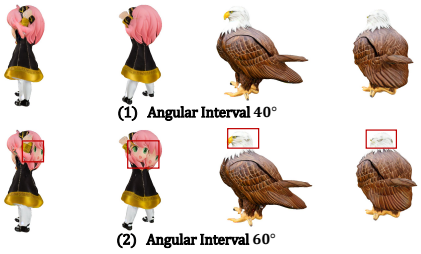


**Fig. 2:** Visual comparison between 40° and 60° as angular interval. Large intervals tend to exhibit multi-face issues.

| **Interval\Metric** | CLIP↑ | Contextual↓ | PSNR↑ | LPIPS↓ |
|---|---|---|---|---|
| 20° | 0.873 | 1.504 | 22.35 | 0.051 |
| 40° | 0.881 | 1.506 | **22.38** | **0.048** |
| 60° | **0.888** | 1.497 | 22.27 | 0.050 |
| 80° | 0.885 | **1.487** | 22.26 | 0.051 |

**Table 4:** Effects of different angular intervals of camera views on Test-alpha dataset. Large angular interval performs better quantitatively.

## 4 DDNM Option

As shown in Figure 3, the issues in view 1 (120°) stem from repainting the previous view (80°). The diffusion model repaints unseen areas based on what it deems reasonable, but this process can sometimes result in artifacts when viewed from other angles. To address this, we can simply apply image restoration constraints in a zero-shot manner, like DDNM [4], to align the repainted image, after degradation, with the input image. The input image used for repainting, though low-quality, is multi-view consistent because it is optimized via 3D diffusion SDS in the coarse stage. Therefore, this alignment ensures multi-view plausibility. However, while DDNM improves faithfulness, it may reduce realism, as shown in Figure 3. Users can adjust the DDNM strength to achieve the desired trade-off.

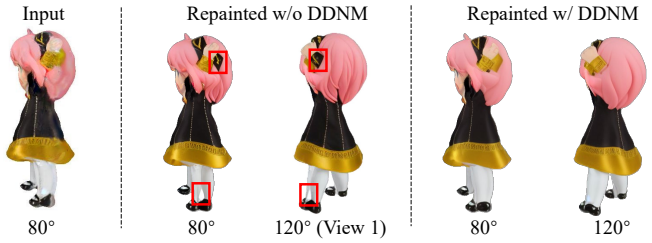


**Fig. 3:** Visualization of multi-view consistency with DDNM constraints.

## 5 More Results

**Visualization of More Repainted Views.** Figure 4 shows one example with all 8 repainted views to demonstrate the multi-view consistency of our method.

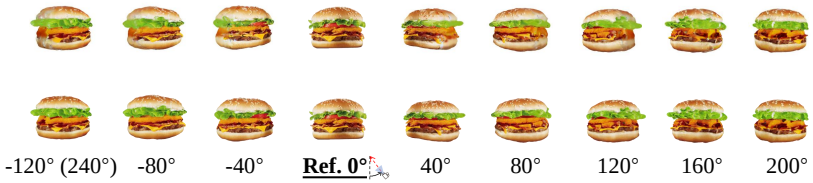


**Fig. 4:** Visualization of all the repainted views. **Top line:** before repainting. **Bottom line:** after repainting. *Zoom in for details.*

**Results on DTU and NeRF4 datasets.** As shown in Figure 5, we conducted experiments on NeRF4 [3] and DTU [2] datasets. Our results demonstrate better consistency and finer textures compared to DreamGaussian and Magic123.


**Fig. 5:** Visual comparison on DTU and NeRF4 datasets.

## References

1. Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google scanned objects: A high-quality dataset of 3d scanned household items. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2553–2560. IEEE (2022)
2. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanaes, H.: Large scale multi-view stereopsis evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2014)
3. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
4. Wang, Y., Yu, J., Zhang, J.: Zero-shot image restoration using denoising diffusion null-space model. The Eleventh International Conference on Learning Representations (2023)