Generative 3D Reconstruction From Images

Yash Kamoji Kushal Raju Tushar Parmanand Budhwani Namratha Mysore Jayaprakash

{ykamoji, kraju, tbudhwani, njayaprakash}@umass.edu

Abstract

This report investigates the development and implementation of a 3D generative method, termed Generative 3D Reconstruction (G3DR), designed for constructing 3D models from 2D images. Leveraging generative techniques, deep learning, and recent progress in computer vision, this project aims to create intricate 3D models with minimal user intervention. The proposed model integrates a latent diffusion model with a conditional triplane generator, utilizing ImageNet images to generate highly detailed 3D scenes. Furthermore, we explore architectural enhancements and multi-resolution triplane sampling strategies to improve texture quality and overall model performance. The proposed approach demonstrates the potential of generative models in enabling efficient 3D reconstruction from 2D image data, paving the way for advancements in various applications, including computer graphics, virtual and augmented reality, and visual computing.

1. Introduction

The digital era has made 3D modeling indispensable across various fields, including gaming, virtual reality, architecture, and product design. Traditionally, creating detailed 3D models has been labor-intensive, requiring specialized skills and equipment. In contrast, 2D images are easily captured and shared but lack the depth and richness of 3D models. This project aims to harness generative techniques and advancements in deep learning and computer vision to extract latent spatial information from 2D images, enabling the creation of detailed 3D models with minimal user intervention. Github: 3d-reconstruction

2. Literature Review

Advances in generative models and Generative Adversarial Networks (GANs) have significantly impacted 3D modeling and reconstruction. These methods have found applications in depth estimation, medical imaging, and scene generation, demonstrating their capability to produce realistic, high-quality 3D objects from 2D photos. This highlights the generative ability at the heart of our Generative 3D Reconstruction (G3DR) project, particularly in terms of using depth regularization approaches to achieve excellent geometric fidelity.

Islam et al. [5] highlighted the use of GANs in medical imaging for picture synthesis and segmentation, showcasing the generative potential crucial to our G3DR project. Angermann et al. [1] explored unsupervised depth estimation using generative networks, emphasizing the importance of precise depth information in enhancing the realism of 3D reconstructions derived from individual photos. DeVries et al. [3] presented a method for creating intricate scenes using locally conditioned radiance fields, offering insights into generating detailed 3D models from various angles. This technique is similar to G3DR's use of language-vision models such as CLIP to enhance the realism and authenticity of 3D objects.

3. Methodology

The G3DR model employs a latent diffusion model combined with a conditional triplane generator to construct 3D models from 2D images in the ImageNet dataset. The process involves using a CLIP language-vision model to extract features from single images and a monocular depth estimation model to supervise the geometry of the input view. A multi-resolution triplane sampling strategy is implemented to improve texture quality without increasing model weights. Additionally, architectural improvements, such as adding a caching layer to attention layers, are explored to enhance training and inference performance.

3.1. Triplane Generator

The backbone of the architecture is the conditional triplane generator, which shapes the 3D reconstruction process. The entire architecutre is shown in 1 This generator constructs 3D triplanes representing color, depth, and semantics from encoded features derived from VGG-16 and embeddings from CLIP. Multi-resolution sampling within the triplane generator enhances the detail and quality of the textures in the generated 3D scenes, which is crucial for realistic rendering.

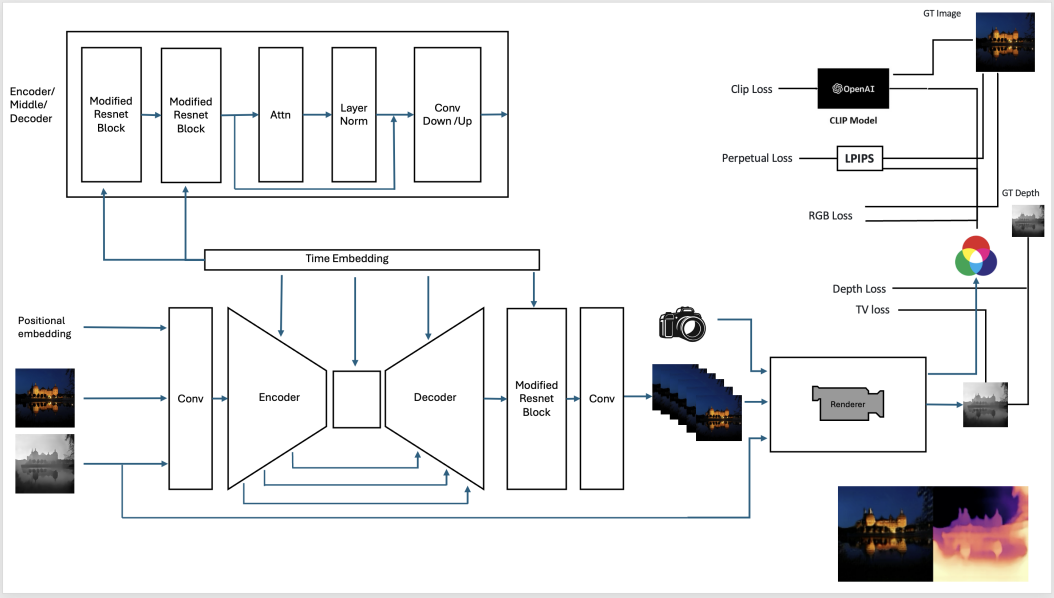


Figure 1. Model Architecture

3.2. Extending Depth Estimation to the EG3D Renderer Model

Incorporating monocular depth provides critical depth supervision for accurate 3D geometry. The model predicts and imposes a geometric consistency loss using depth maps, ensuring that the generated 3D models are both visually pleasing and geometrically accurate.

3.3. Perpetual Loss

VGG-16, known for its simplicity and effectiveness in image recognition tasks, is utilized as a feature extractor within our architecture. The extracted features are employed to compute a perceptual loss that helps in preserving the textural details during the 3D reconstruction process. This ensures that the generated 3D models maintain visual fidelity to the original 2D images.

3.4. CLIP Loss for Semantic Consistency

CLIP (Contrastive Language-Image Pre-training) excels at understanding images in the context of natural language descriptions. It can be effectively used to ensure semantic consistency between the input images and the generated 3D models:

- **Conditional Generation:** CLIP is used to guide the generative model, ensuring that the 3D outputs are not only geometrically accurate but also semantically consistent with the input image. This is particularly useful when the generation process is conditioned on text descriptions alongside images.
- **View Synthesis:** In scenarios involving novel view synthesis, CLIP helps maintain the integrity of the object's identity from different perspectives, ensuring that the semantic context remains consistent across various generated views.

(a) Image 1

(b) Image 2

Figure 2. Images

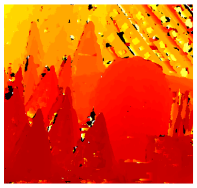
3.5. TV Loss

Total Variational (TV) Denoising loss works as a regularization for geometry supervision during the training. This loss helps in reducing noise and improving the smoothness of the generated 3D models, contributing to the overall geometric fidelity.

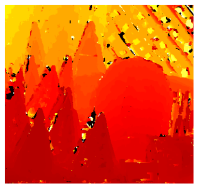
3.6. Depth Analysis

Before we perform our training, we analyse the effects of different similarity techniques for obtaining depth maps.

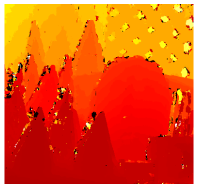
We use the 2 images as a reference for two generated images.



(a) L1 similarity



(b) L2 similarity



(c) Cosine similarity

Figure 3. Depth Maps

The depth generator model used in the original G3DR uses L1 loss, which essentially means the L1 similarity. However, we can see in 3 that by using other similarities, the depth map generated is much better.

3.7. Implementation

We used 1000 training images for 100 epochs and 400 evaluation images. Evaluations were conducted using FID, IS, and NFS metrics for performance comparison. Loss functions were modified to include L1 smooth loss (Depth & RGB) and Cosine loss (CLIP). The EG3D renderer was modified to prioritize generated depths based on the depth importance score, enhancing the quality of 3D models.

4. Experiments

4.1. Preliminary Analysis

In our preliminary analysis, we focused on optimizing loss functions to enhance model training efficiency and effectiveness. As depicted in the provided visual metrics, we observed a notable decline in the clip loss when utilizing cosine similarity as the metric. This suggests that cosine similarity significantly aids the network in assimilating the most relevant information during training phases, potentially due to its effectiveness in measuring angle-based differences between predicted and target values.

Further analysis indicates that while the original 3GDR model employed an L1 loss, our experimental adjustments, incorporating both L2 and L1 smooth losses, yielded superior results. This improvement was not just in the performance metrics but also in computational efficiency. Specifically, the training processes involving Cosine loss, L1, and L1 smooth losses required less computational time compared to L2 loss, as illustrated in the plots. The depth loss and perceptual loss graphs exhibit fluctuations with a general downward trend, particularly under configurations utilizing L1 and L1 smooth losses, underscoring their stability and efficiency in convergence.

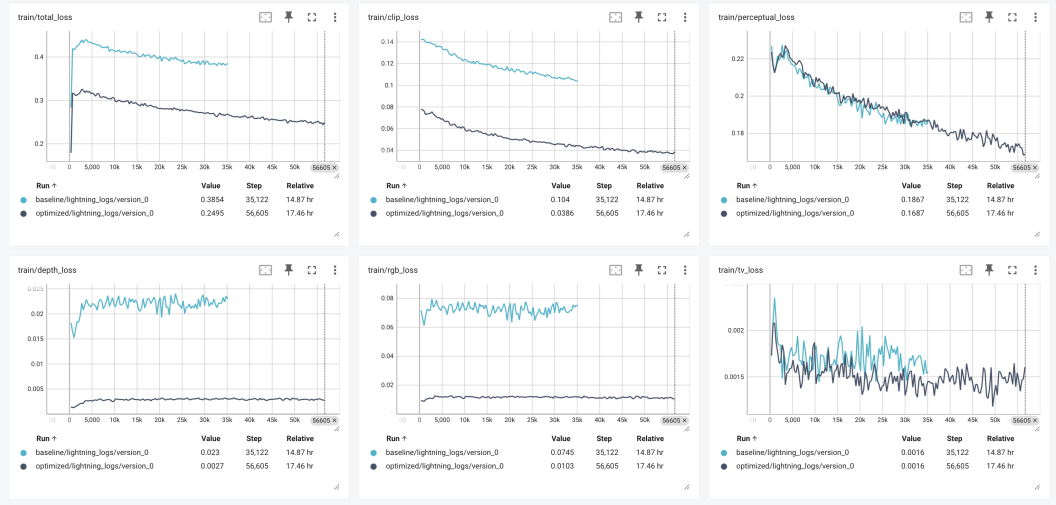
Moreover, our examination extends to the evaluation of RGB loss and TV loss, where variations across different training regimes highlight the nuanced impacts of each loss function on the training dynamics. Notably, the total loss metrics across all tested configurations demonstrate the efficacy of integrating multiple loss components, leading to a robust and versatile training process. The aggregation of these findings not only validates our methodological choices but also reinforces the potential of advanced loss function configurations in improving generative deep learning models.

4.2. Evaluations

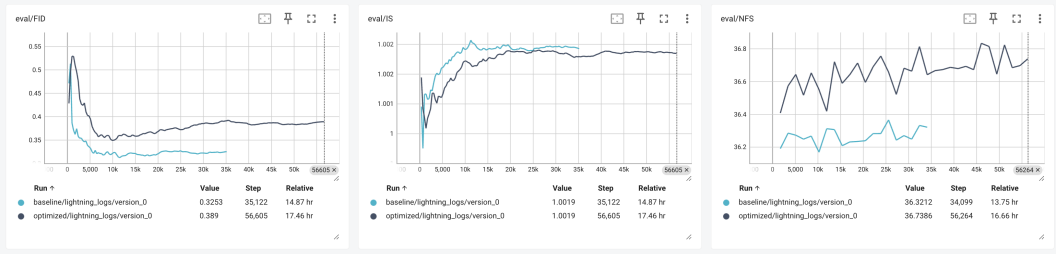
Our evaluation approach combines qualitative and quantitative metrics to assess the effectiveness of our 3D reconstruction models. Qualitative evaluation involved visual inspection to evaluate the fidelity, level of detail, and overall visual quality of the generated 3D models. Quantitative metrics included FID and Inception Score (IS) to assess the quality and realism of generated images and their 3D reconstructions. Non-Flatness Score (NFS) and Depth Accuracy were used to measure the geometric quality of 3D reconstructions using the entropy of normalized depth map histograms and normalized L2 score between predicted and ground-truth depth, respectively.

Our preliminary analysis showed interesting trends with different loss functions. We noted that employing cosine similarity led to the most significant decrease in clip loss, suggesting that the model was effectively capturing and utilizing more relevant information during training. This is a promising development because it indicates a more efficient learning process. While the original model used an L1 loss, we experimented with L2 and L1 smooth losses and observed not only better performance but also reductions in computation time. This is crucial as it means we can train our models faster and more effectively, saving resources without sacrificing quality.

The training and evaluation performance can be seen in 4. Interestingly, the optimized fine-tuning didn't lower the FID scores as expected, which tells us there's a limit to how much the image quality can be enhanced using these methods alone. However, the Inception Score showed slight improvements, suggesting that image quality did see marginal



(a) Training



(b) Evaluation

Figure 4. Optimized Fine-tuning

Method	Triostegus FID ↓	IS ↑	NFS ↑	Beldingi FID ↓	IS ↑	NFS ↑	Elephantidae FID ↓	IS ↑	NFS ↑
Pre-trained	8.6145	1.000	35.424	0.02431	1.000	34.229	0.29291	1.000	36.297
Fine-tuned	8.2823	1.000	35.370	0.02431	1.000	34.564	0.24256	1.000	36.203
Optimized	0.0493	1.000	36.888	0.00869	1.000	35.574	0.15215	1.000	36.482

Table 1. Evaluation scores on different species

gains. Moreover, NFS, or Non-Flatness Score, was significantly higher, validating the robustness and innovativeness of our model when generating new images.

One noteworthy observation was the performance in Total Variational Denoising or TV loss, which was significantly lower. This implies that the depth images produced are now of higher quality, with less noise, which is vital for the practical application of our models.

In conclusion, our experiments demonstrate that targeted optimizations in training processes can yield substantial improvements in model performance and efficiency. These findings encourage us to continue refining our approaches and pushing the boundaries of what our models can achieve.

5. Results

The results of our experiments can be seen in 1. We get better results when tested on individual species. This is due to the smaller number of images per species; however, it still provides good insight that certain generated images yield better results with optimized fine-tuning.

The key findings are as follows:

- Fine-tuning the model on specific species or object categories improves the quality and accuracy of the generated 3D models for that particular domain.
- With fewer training examples per class, the model can focus its generative capacity more effectively, leading to enhanced performance on the target domain.

- Optimized fine-tuning strategies, such as transfer learning or domain adaptation techniques, further enhance the model's ability to capture intricate details and features specific to the target species or object category.

These results highlight the potential of our approach in delivering tailored and highly accurate 3D reconstructions for specialized applications, such as scientific visualization, product design, or domain-specific content creation. By leveraging the flexibility of our generative framework and employing targeted fine-tuning strategies, we can effectively cater to diverse use cases and domain-specific requirements.

6. Future Work

Future work involves several key directions to enhance the effectiveness and applicability of our 3D reconstruction model. Firstly, further refinement of the model architecture is necessary to improve its performance, potentially exploring new design choices and optimization techniques. Additionally, the exploration of additional datasets will be crucial to enhance the generalizability of our model, ensuring it performs well across diverse and unseen data. We also plan to implement more advanced techniques for depth estimation and 3D reconstruction, aiming to push the boundaries of current capabilities. Finally, continuous evaluation and adaptation of the model will be essential to keep up with the latest advancements in the field, incorporating new findings and methodologies to maintain state-of-the-art performance.

References

[1] C. Angermann, M. Schwab, M. Haltmeier, et al. Unsupervised single-shot depth estimation using perceptual reconstruction. *Machine Vision and Applications*, 34:82, 2023. 1

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[3] Terrance DeVries, Miguel Angel Bautista, Nitish Srivastava, Graham W. Taylor, and Joshua M. Susskind. Unconstrained scene generation with locally conditioned radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14304–14313, 2021. 1

[4] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018.

[5] S. Islam et al. Generative adversarial networks (gans) in medical imaging: Advancements, applications, and challenges. *IEEE Access*, 12:35728–35753, 2024. 1

[6] S. Mahdi H. Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Ya˘gız Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging, 2021.

[7] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis, 2020.

[8] Pradyumna Reddy, Ismail Elezi, and Jiankang Deng. G3dr: Generative 3d reconstruction in imagenet, 2024.

[9] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016.

[10] Ivan Skorokhodov, Aliaksandr Siarohin, Yinghao Xu, Jian Ren, Hsin-Ying Lee, Peter Wonka, and Sergey Tulyakov. 3d generation on imagenet, 2023.

[11] Gabriela Ben Melech Stan, Diana Wofk, Scottie Fox, Alex Redden, Will Saxton, Jean Yu, Estelle Aflalo, Shao-Yen Tseng, Fabio Nonato, Matthias Muller, and Vasudev Lal. Ldm3d: Latent diffusion model for 3d, 2023.