Autodecoder Latent Space 3D Diffusion

Akarsh Aurora
Massachusetts Institute of Technology
77 Massachusetts Ave, Cambridge, MA 02139
aaurora@mit.edu

1. Problem

In the rapidly growing domain of 3D content generation for video games, mixed reality, and entertainment, efficiently representing 3D assets has become a topic of increasing research interest [1]. Implicit neural representations (INR) have gained prominence among representation techniques due to their continuous, differentiable nature, facilitating practical applications such as style transfer and real-time editing [2]. While INRs are expressive and flexible, they traditionally necessitate costly optimization for each dataset sample. Techniques like DreamFusion, which conditions on images [3], and related methodologies [4], as well as 3D autoencoder models with implicit decoders [5], typically require either extensive computational resources per generation or substantial amounts of 3D data. This makes them less viable for consumer-grade applications in a context where computing resources are low and high-quality generations are necessary.

To address these challenges, the autodecoder framework emerges as a promising solution [6]. Distinguished by its use of 2D supervised meta-learning for parameterizing 3D latent spaces, autodecoders effectively minimize the typical reliance on 3D data availability. Notably, when focusing on singular object classes, autodecoders demonstrate a marked reduction in computational demand and capture intricate data distributions efficiently. In our approach, we leverage the autodecoder's capabilities to diffuse a 3D latent space, learned with image supervision of ShapeNet objects [7]. We hypothesize that this method will maintain the rapid inference capabilities of SOTA 3D diffusion models while simultaneously mitigating the need for extensive 3D data, which is otherwise scarcely available in the wild.

2. Related Work

Prior work with 3D generation has centered on auto-encoders trained on explicit 3D representations, with subsequent generative modeling in the resultant latent spaces. Researchers have experimented with various generative models, including GANs [8] and normalizing flows [9], to model these spaces, and employed diffusion models [10][11] for enhanced depth and complexity in generation. Nonetheless, these fixed, explicit representations often encounter issues with resolution and are limited in their ability to create smoothly rendered 3D assets.

In parallel, there has been a push towards implicit decoders within 3D auto-encoders. Techniques involving encoding differentiable signed distance field (SDF) samples [12] or voxel grids [13] into latent spaces that condition implicit models have been explored, with some approaches directly producing parameters for multi-layer perceptions (MLP) conditioned on rendered views using transformer-based architectures [14][15]. Strategies to create latent-conditional implicit 3D representations without a learned encoder have also emerged. These endeavors have laid the groundwork for more expressive implicit representations and scalable training processes. However, these methods face scalability challenges, as each new instance requires multiple gradient steps, making large dataset applications compute-intensive. Additionally, the effectiveness of both implicit and explicit auto-encoding approaches rests heavily on the amount of 3D data used for training.

Recent advances have also seen the integration of large-scale, pre-trained 2D diffusion models into the realm of 3D generation, as evidenced by several studies [3][16][17]. The innovative aspect of these methods lies in utilizing 2D diffusion models to assess renderings from diverse viewpoints. This assessment is then utilized to refine a 3D-aware representation of the content. Although effective, these approaches often entail a resource-intensive optimization process for generating each new object due to their inherent reliance on depth estimation. In contrast to these proposed methods, our approach leverages a more intuitive and efficient 3D representation strategy that reduces the computational burden associated with traditional 3D generation strategies.

3. Method

In our methodology, we propose a two-step approach as illustrated in Fig. 1. Initially, we train an autodecoder encompassing a collection of embedding vectors. These vectors correspond to objects in the training dataset and are utilized to create a low-resolution latent 3D feature volume. This volume undergoes progressive upsampling before being decoded into a voxelized form, representing the shape and appearance of the generated object. The network is trained using 2D reconstruction supervision from training images and volumetric rendering techniques.

After training, the autodecoder is split into two components: the 3D convolutional decoding upsampler (G1) and the radiance field renderer (G2). This separation enables the training of a 3D diffusion model on the learned latent 3D space derived from the autodecoder. The diffusion process leverages the structural and appearance attributes acquired during autodecoder training, enabling the generation of diverse and realistic 3D content with relatively high efficiency.

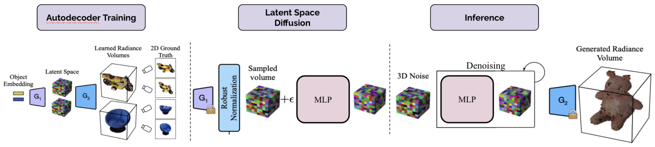

Figure 1. In the training process, an autodecoder uses two components, G1 and G2. G1 converts each object in the training set into a latent volumetric space, while G2 then decodes these volumes into radiance volumes for image rendering. This is achieved with only 2D supervision. Once the autodecoder is trained, its weight parameters are frozen. Then, G1's latent volumes are used to train a 3D denoising diffusion process. During inference, volumes are randomly generated and denoised, then decoded by G2 for rendering. Figure adapted from Evangelos *et al*. [18].

3.1. Autodecoder Architecture

At the heart of the autodecoder is the concept of a 3D latent space, which provides a compact and rich representation of the scene. The 3D latent space is embodied in a feature grid, a learnable tensor that encodes spatial information. This grid is initialized with small random values or latent codes that serve as the foundational representation of the scene. Depending on the mode of operation, this grid can represent spatial information in either 2D or 3D, catering to different levels of complexity in scene representation.

The decoding process involves a convolutional network, which takes the encoded feature grid and progressively decodes it into a detailed feature map. To enrich the representation with spatial context, a position concatenation mechanism is employed. This mechanism dynamically generates a grid of coordinates, which are then concatenated with the feature map, thus integrating spatial information directly into the feature representation.

The neural radiance field renderer, another critical component of the architecture, maps 3D coordinates to color and density values, essential for the final image rendering. This renderer combines a grid-based feature representation with a multi-layer perceptron (MLP), allowing for efficient encoding and decoding of complex scenes. The rendering process is performed through a volume rendering technique, which involves marching rays through the scene and accumulating color and density values along these rays. This method produces smooth images by accurately simulating the way light interacts with the scene.

For training and optimization, we employ a custom loss function that balances image reconstruction accuracy with regularization terms to ensure the stability of the learning process. The entire system is trained end-to-end, allowing the model to learn a coherent representation of the scene and how to render it effectively.

3.2. Latent Space 3D Diffusion Architecture

The 3D diffusion model is specifically designed to operate within the learned latent 3D space of the autodecoder. The process is conceptualized as a Markov chain, progressively adding Gaussian noise over a predefined number of timesteps to normalized, sampled latent 3D feature volumes. Initially, the latent space captures rich semantic information about the 3D objects. As the diffusion steps progress, this information is increasingly corrupted by noise, effectively transforming the data into a Gaussian random field by the final step.

Using an MLP, our model learns to reverse the stochastic noising process and generate out-of-distribution 3D latents [21]. The denoising is conditioned on both the current state of the latent representation and the specific timestep via concatenation, with the latter being encoded through sinusoidal time embeddings to help learn high-frequency functions [22]. The predicted, denoised latent volumes are then fed into the radiance field renderer to produce a new radiance volume, resulting in coherent 3D content.

3.3. ShapeNet Dataset

The model is trained on the ShapeNet earphone subset containing 72 distinct objects. We use a fixed lighting configuration that only supports ambient shading and render 30 views of each instance at a resolution of 128 × 128 pixels. Camera poses are randomly generated for the image renderings by varying the elevation and azimuth of the pose while keeping the object at the origin. Before entering the autodecoder training loop, images were clipped and downsampled to a given side length for computational efficiency. We evaluate qualitative performance on the reconstructed view fidelity of objects in the training set and the novel-view fidelity of generated objects from latent 3D diffusion.

4. Experiments

In our initial experiment, we rendered the ShapeNet earphones using PyTorch3D and converted the extrinsics into the OpenCV camera convention using a custom conversion script. Additionally, we reimplemented positional embeddings, a convolutional decoder, a hybrid voxel neural field grid, a radiance field approximator, and a volume renderer. These elements collectively facilitated scene fitting for an individual earphone model. Post 1,000 training iterations, the fitting outcome is displayed in Fig. 2. Notably, while the trained MLP exhibited limitations in color differentiation, the depth map's precision underscored the effectiveness of the convolutional autodecoder.

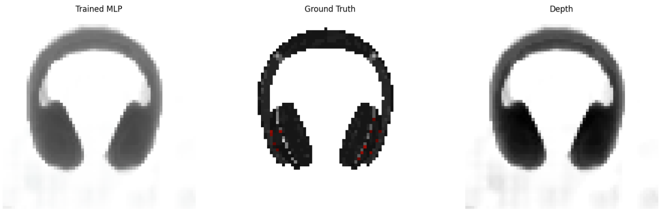

Figure 2. Trained MLP output, ground truth input, and depth map output of single-scene fitting with convolutional autodecoder.

For our concluding experiment, we developed a multi-object autodecoder wrapper and extended our training to encompass the entire set of earphone models from ShapeNet. Through this approach, we generated and denoised random samples from the latent 3D space, subsequently processing them via our rendering pipeline. This procedure yielded novel earphone designs, as depicted in Fig. 3. The resultant 3D model, not present in the training dataset, convincingly resembles an actual earphone. A key observation is the distinct separation between the ear cup and the top rim, demonstrating the diffusion model's capacity to adhere to the characteristic features of the earphone category while still generating novel volumetric forms.

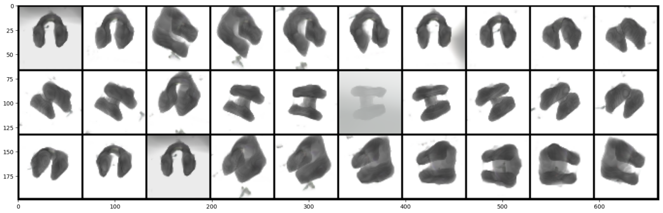

Figure 3. Varied camera pose images of a novel 3D headphone generated by autodecoder latent space 3D diffusion pipeline.

5. Limitations and Future Work

Our current autodecoder framework, while showing promise in generating 3D objects, faces several key limitations. A significant challenge is the dependency on extensive training within the latent space to achieve high-quality outputs. As this study utilized a low-compute version of the framework, it primarily resulted in the generation of basic objects with limited fidelity. Additionally, the framework, in its current incarnation, is primarily tailored to non-complex, single-object scenes. It does not inherently accommodate the complexities arising from the interactions of multiple distinct objects, such as occlusions or spatial arrangements. This is affirmed by previous work, which also find that autodecoder latent codes are too constrained to represent multi-object scenes [18][20].

In our future work, we aim to extend the training regime and include a more varied array of object categories, especially those with uncommon features. This expansion is anticipated to test and potentially amplify the framework's ability to understand and generate a more diverse set of objects, thereby broadening its applicability in practical 3D generation scenarios. Another promising direction involves the incorporation of self-attention mechanisms within the autodecoder to harmonize the appearance of the reconstructed objects for enhanced color discrimination and texture. Furthermore, the adoption of U-Net architectures in place of the current MLP for diffusion would enable segmentation capabilities, which are crucial for generating complex objects with greater spatial accuracy [23].

An especially exciting avenue for future exploration is the integration of CLIP (Contrastive Language-Image Pretraining) [24] models with our diffusion processes. This integration could enable the creation of content that is responsive to natural language inputs, aligning the outputs more closely with human-centric descriptions and on-demand generation needs. In summary, while our current framework lays a solid groundwork in the realm of 3D object generation, these prospective developments aim to significantly elevate its capabilities to handle more complex scenarios and respond to nuanced generative demands.

6. Conclusion

Our work demonstrates the feasibility of using a 2D supervised model to generate novel 3D assets through an autodecoder latent space diffusion approach. This method leverages the autodecoder's unique ability to synthesize content without direct encoded 3D inputs, learning representations that capture the structure and appearance of basic objects. Such capability is crucial for producing 3D objects under the constraints of 2D supervision. Employing a latent volumetric representation, our model facilitates effective 3D diffusion modeling and view-consistent rendering. Our results are particularly notable given the small, single-category, synthetic dataset and computational resources used, highlighting the potential of autodecoders in helping democratize 3D asset generation. This work points towards new avenues in 3D content creation, emphasizing the possibility of achieving high-fidelity outputs with minimal 3D data and low compute, setting the stage for new advancements in accessible and efficient 3D modeling techniques.

References

[1] A. Jain, B. Mildenhall, J. T. Barron, P. Abbeel, and B. Poole, ”Zero-Shot Text-Guided Object Generation with Dream Fields,” *arXiv preprint arXiv:2112.01455*, May 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2112.01455

[2] C. Bao *et al.*, ”SINE: Semantic-driven Image-based NeRF Editing with Prior-guided Editing Field,” *arXiv preprint arXiv:2303.13277*, Mar. 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.13277

[3] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, ”DreamFusion: Text-to-3D using 2D Diffusion,” *arXiv preprint arXiv:2209.14988*, Sep. 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2209.14988

[4] C.-H. Lin *et al.*, ”Magic3D: High-Resolution Text-to-3D Content Creation,” *arXiv preprint arXiv:2211.10440*, Mar. 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2211.10440

[5] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, ”DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation,” *arXiv preprint arXiv:1901.05103*, Jan. 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1901.05103

[6] Y. Xie, T. Takikawa, S. Saito, O. Litany, S. Yan, N. Khan, F. Tombari, J. Tompkin, V. Sitzmann, and S. Sridhar, ”Neural Fields in Visual Computing and Beyond,” *CoRR*, vol. abs/2111.11426, 2021. [Online]. Available: https://arxiv.org/abs/2111.11426

[7] A. X. Chang *et al.*, ”ShapeNet: An Information-Rich 3D Model Repository,” *arXiv preprint arXiv:1512.03012*, Dec. 2015. [Online]. Available: https://doi.org/10.48550/arXiv.1512.03012

[8] I. J. Goodfellow *et al.*, ”Generative Adversarial Networks,” *arXiv preprint arXiv:1406.2661*, Jun. 2014. [Online]. Available: https://doi.org/10.48550/arXiv.1406.2661

[9] D. J. Rezende and S. Mohamed, ”Variational Inference with Normalizing Flows,” *arXiv preprint arXiv:1505.05770*, Jun. 2016. [Online]. Available: https://doi.org/10.48550/arXiv.1505.05770

[10] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, ”Deep Unsupervised Learning using Nonequilibrium Thermodynamics,” *arXiv preprint arXiv:1503.03585*, Nov. 2015. [Online]. Available: https://doi.org/10.48550/arXiv.1503.03585

[11] E. Dupont, H. Kim, S. M. A. Eslami, D. Rezende, and D. Rosenbaum, ”From data to functa: Your data point is a function and you can treat it like one,” *arXiv preprint arXiv:2201.12204*, Nov. 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2201.12204

[12] R. Fu, X. Zhan, Y. Chen, D. Ritchie, and S. Sridhar, ”ShapeCrafter: A Recursive Text-Conditioned 3D Shape Generation Model,” *arXiv preprint arXiv:2207.09446*, Apr. 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2207.09446

[13] A. Sanghi *et al.*, ”CLIP-Forge: Towards Zero-Shot Text-to-Shape Generation,” *arXiv preprint arXiv:2110.02624*, Apr. 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2110.02624

[14] A. Vaswani *et al.*, ”Attention Is All You Need,” *arXiv preprint arXiv:1706.03762*, Aug. 2023. [Online]. Available: https://doi.org/10.48550/arXiv.1706.03762

[15] Y. Chen and X. Wang, ”Transformers as Meta-Learners for Implicit Neural Representations,” *arXiv preprint arXiv:2208.02801*, Aug. 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2208.02801

[16] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin, ”Magic3D: High-Resolution Text-to-3D Content Creation,” *arXiv preprint arXiv:2211.10440*, 2023. [Online]. Available: https://arxiv.org/abs/2211.10440

[17] R. Chen, Y. Chen, N. Jiao, and K. Jia, ”Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation,” *arXiv preprint arXiv:2303.13873*, 2023. [Online]. Available: https://arxiv.org/abs/2303.13873

[18] E. Ntavelis, A. Siarohin, K. Olszewski, C. Wang, L. Van Gool, and S. Tulyakov, ”AutoDecoding Latent 3D Diffusion Models,” *arXiv preprint arXiv:2307.05445*, 2023. [Online]. Available: https://arxiv.org/abs/2307.05445

[19] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, ”NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis,” *arXiv preprint arXiv:2003.08934*, Aug. 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2003.08934

[20] V. Sitzmann, M. Zollhöfer, and G. Wetzstein, ”Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations,” *CoRR*, vol. abs/1906.01618, 2019. [Online]. Available: http://arxiv.org/abs/1906.01618

[21] J. Ho, A. Jain, and P. Abbeel, ”Denoising Diffusion Probabilistic Models,” *CoRR*, vol. abs/2006.11239, 2020. [Online]. Available: https://arxiv.org/abs/2006.11239

[22] M. Tancik, P. P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. T. Barron, and R. Ng, ”Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains,” *CoRR*, vol. abs/2006.10739, 2020. [Online]. Available: https://arxiv.org/abs/2006.10739

[23] C. Si, Z. Huang, Y. Jiang, and Z. Liu, ”FreeU: Free Lunch in Diffusion U-Net,” *arXiv preprint arXiv:2309.11497*, 2023. [Online]. Available: https://arxiv.org/abs/2309.11497

[24] A. Radford *et al.*, ”Learning Transferable Visual Models From Natural Language Supervision,” *arXiv preprint arXiv:2103.00020*, Feb. 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2103.00020