3D-WAG: Hierarchical Wavelet-Guided Autoregressive Generation

Tejaswini Medi [1] Arianna Rampini[2] Pradyumna Reddy[2] Pradeep Kumar Jayaraman[2] Margret Keuper[1,3]

[1]Machine Learning, University of Mannheim [3]Max Planck Institute for Informatics, Saarland Informatics Campus
[2]Autodesk AI Lab

Motivation

- 3D shape generation is crucial for applications like AR/VR and digital design.
- Existing models often struggle with generating high-resolution and coherent 3D shapes efficiently.
- Most hierarchical generative approaches focus on token-level predictions, limiting scalability and geometric detail.
- Can we improve hierarchical generation of high-fidelity 3D Shapes using better framework and representation?

State-of-the-Art

- Diffusion Models (UDiff).
- AR Generation Models (ImAM).
- GANs (V-GAN).

Contribution

- 3D-WAG, a novel wavelet-guided autoregressive framework for high-fidelity 3D shape generation.
- Redefines the generation task as "next-scale" prediction, enabling efficient hierarchical generation using wavelet input representation.

Loss Functions

Stage-1 Autoencoder Loss:

$$L = \lambda_{recon}\|W - \hat{W}\|_2 + \lambda_{commit}\sum_{k=1}^{K}\|z_k^{res} - \hat{z}_k^{res}\|_2.$$

Stage-2 Prior Model Loss:

$$L_{\text{CE}} = -\sum_{i=1}^{K}\log p(f_i|f_{<i}),$$

$p(f_i|f_{<i})$: Probability of predicting the i-th token map f_i given all previous token maps $f_{<i}$ by minimizing the negative log-likelihood.

Summary

- Unlike traditional methods, 3D-WAG leverages multi-scale wavelet representations to preserve global structure and fine-grained details.
- We demonstrate state-of-the-art performance on both unconditional and conditional 3D generation tasks, including text-to-3D.

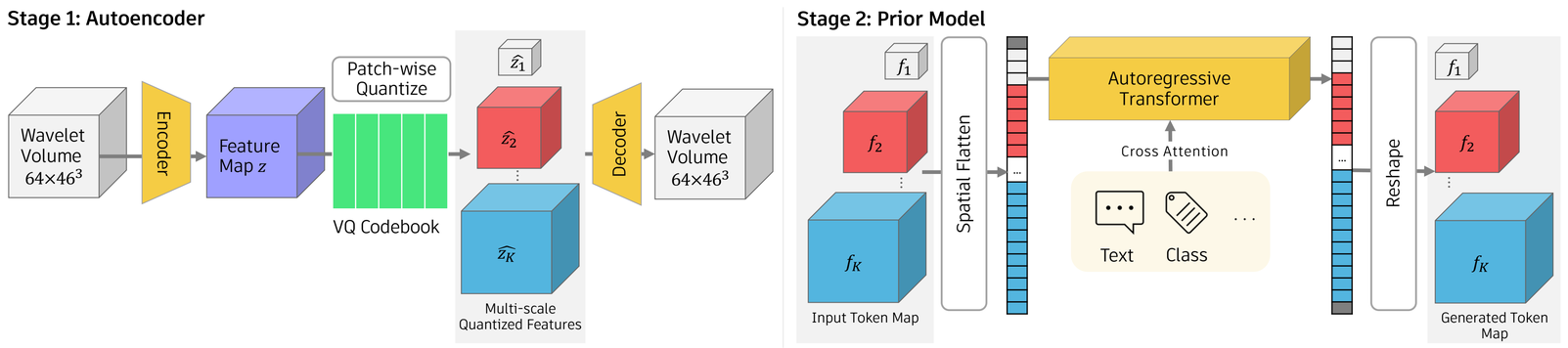
Future Work

- Scaling for large datasets, exploring alternative generative conditioning mechanisms.

3D-WAG Autoregressive Generation

3D-WAG Method

- Our approach is two-stage training approach. Input: Wavelet Volume Data.
- In Stage-1, we train a VQ-VAE with multi-scale patch-wise quantization.
- In Stage-2, codebook indices in the form of token maps are learned with an autoregressive decoder-only Transformer prior model.



Experiments

- Dataset: ShapeNet [1].
- Metrics: Coverage (COV), Minimum Matching Distance (MMD), 1-Nearest Neighbor Accuracy (1-NNA), and Inference Time.
- Highlights: Our model achieves the best COV, MMD, and 1-NNA scores with the fastest inference time.

Comparison of shape generation performance and inference time on the ShapeNet planes and chairs dataset.

Method	Chair						Airplane						Time (s)
	COV ↑		MMD ↓		1-NNA ↓		COV ↑		MMD ↓		1-NNA ↓		
	CD	EMD	CD	EMD	CD	EMD	CD	EMD	CD	EMD	CD	EMD	
IM-GAN	57.01	55.23	12.02	14.76	62.15	63.78	62.21	63.10	3.298	8.479	76.50	76.29	2.98
V-GAN	44.29	39.89	15.35	17.59	80.74	81.62	39.01	39.78	5.875	11.53	93.60	92.98	2.75
PointDiff	51.88	56.11	12.92	16.35	62.02	63.95	60.55	62.71	3.615	9.620	74.98	72.63	–
SPAGHETI	50.11	50.73	15.02	16.05	72.82	69.93	57.30	59.12	4.342	9.021	79.84	79.33	27.4
SALAD (Global)	50.13	49.12	11.93	14.35	63.25	61.73	55.42	59.85	3.891	9.073	82.61	80.87	33.2
SALAD	56.91	55.65	11.89	14.50	58.23	58.90	63.65	65.82	3.721	8.329	74.35	71.52	38.0
WaveGen	50.12	50.67	12.31	14.47	65.58	63.39	61.40	59.57	3.605	8.075	76.23	73.25	15.4
ImAM	54.38	55.71	11.90	13.97	62.43	62.11	65.82	65.41	3.248	7.900	77.12	79.13	14.2
UDiff	53.02	56.51	11.82	14.19	66.42	63.91	65.23	64.22	3.229	7.879	74.85	79.42	39.0
Ours	57.30	58.41	10.90	13.06	68.10	66.31	66.80	65.93	2.900	7.543	82.10	80.92	2.50

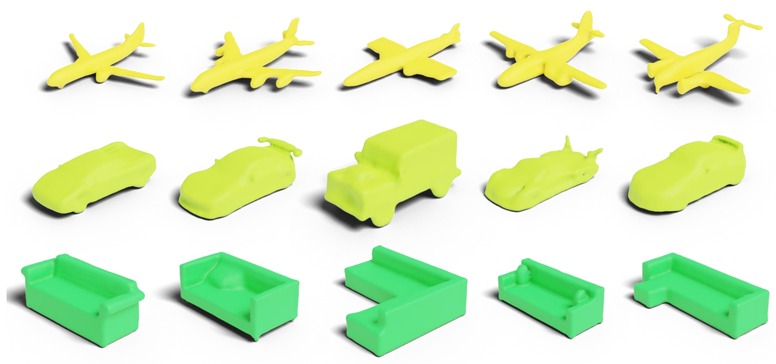
Experiments

- Dataset: DeepFashion3D [2].
- Metrics: Coverage (COV), Minimum Matching Distance (MMD), (1-NNA).
- MMD-CD and MMD-EMD scores are scaled by 10^3 and 10^2.

Comparison of shape generation on the DeepFashion3D dataset.

Method	COV ↑		MMD ↓		1-NNA ↓	
	CD	EMD	CD	EMD	CD	EMD
PointDiff	68.90	64.89	11.20	15.72	83.50	87.93
WaveGen	62.75	52.15	15.78	17.29	93.20	94.96
Diff-SDF	67.45	62.40	14.92	16.88	89.31	92.81
LAS-Diffusion	67.82	56.43	14.75	16.79	88.95	91.68
UDiff	69.98	68.01	11.75	14.25	82.11	82.47
Ours	69.85	68.03	10.96	13.89	82.01	82.30

Conditional Generation



Examples of shape generation conditioned by class: airplane (top), car (middle), and sofa (bottom) categories.

Examples of our text-conditioned shape generation. The results show our approach generating globally coherent output that follows the input text condition.

References

[1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3D model repository. 2015.

[2] Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images.