Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.079254

ARTICLE

SWAGE-3D: Spectral Wasserstein Attention Generative Ensemble, A Comparative Analysis on the ShapeNet Dataset

Zafer Serin[1,*], **Cihan Karakuzu**[2] **and Uğur Yüzgeç**[2]

[1]Pazaryeri Vocational School, Bilecik Seyh Edebali University, Bilecik, Türkiye
[2]Department of Computer Engineering, Bilecik Seyh Edebali University, Bilecik, Türkiye

*Corresponding Author: Zafer Serin. Email: zafer.serin@bilecik.edu.tr

Received: 18 January 2026; Accepted: 20 March 2026; Published: 27 May 2026

ABSTRACT: This study proposes SWAGE-3D (Spectral Wasserstein Attention Generative Ensemble), an enhanced 3D-VAE-GAN framework for single-view 3D object reconstruction using voxel-based representations. The proposed model integrates RGB-D encoding, Wasserstein adversarial learning with hybrid Lipschitz regularization, and a self-attention–augmented generator to improve structural coherence and training stability. By combining variational latent modeling with stabilized Wasserstein optimization, the framework aims to address common challenges in 3D generative modeling, including mode collapse, unstable convergence, and insufficient global consistency. The encoder employs a depth-aware feature extraction strategy, while the discriminator utilizes a hybrid spectral normalization and gradient penalty mechanism to ensure robust approximation of the Wasserstein objective. Additionally, an ensemble strategy is applied at inference time to enhance reconstruction reliability. The proposed approach is evaluated on the ShapeNet dataset across 13 object categories using the Intersection over Union (IoU) metric. Experimental results demonstrate a 16.7% improvement over the baseline 3D-VAE-GAN and competitive performance against state-of-the-art voxel-based reconstruction methods. These findings confirm that the synergistic integration of depth cues, stabilized Wasserstein training, and attention mechanisms significantly enhances single-view 3D reconstruction performance.

KEYWORDS: Three-dimensional reconstruction; variational autoencoder; generative adversarial network; depth estimation; residual neural network; ensemble learning; attention mechanism

1 Introduction

Single-view 3D object reconstruction remains a fundamental challenge in computer vision due to the inherently ill-posed nature of inferring volumetric geometry from a single 2D observation. The problem has broad applications in virtual reality, robotics, and digital content creation. Traditional manual 3D modeling is time-consuming and often requires substantial expert effort for producing a single high-quality model [1]. Consequently, learning-based reconstruction methods have emerged as scalable alternatives for accelerating 3D content generation.

Among various 3D representations, voxel-based models provide a structured volumetric formulation that integrates naturally with convolutional neural networks. As illustrated in Fig. 1, voxel grids encode occupancy information within a regular 3D lattice, enabling direct application of 3D convolutions and probabilistic generative modeling [2]. In contrast, point-cloud representations lack explicit surface connectivity and often require post-processing for surface reconstruction [3], while polygon mesh representations demand predefined topology and may suffer from discretization artifacts [4]. Due to their structural

regularity and compatibility with convolutional architectures, voxel representations remain a practical choice for end-to-end reconstruction pipelines.

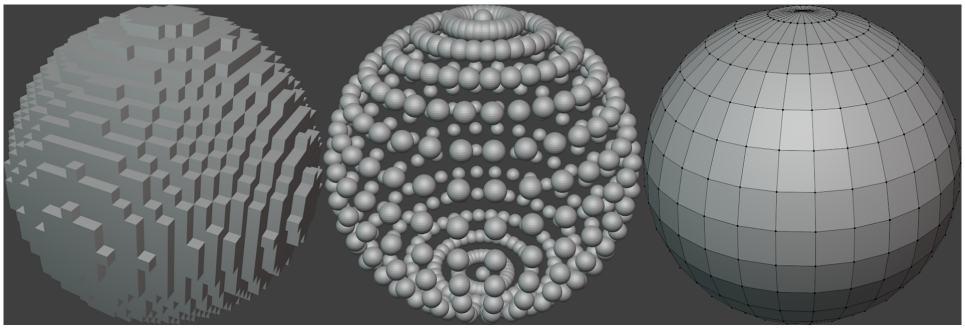

Figure 1: Voxel, point-cloud, and polygon (mesh) representations of a sphere.

Deep learning has substantially advanced image-based feature extraction and generative modeling. Convolutional Neural Networks (CNNs), including architectures such as ResNet [5–8], have demonstrated strong capability in visual representation learning. The inclusion of diverse training strategies and augmentation techniques, such as MixUp, has further improved generalization performance [9,10].

Generative frameworks, particularly Variational Autoencoders (VAEs) [11] and Generative Adversarial Networks (GANs) [12], have enabled probabilistic modeling of volumetric data. However, standard GAN training is prone to instability and mode collapse, especially in high-dimensional voxel spaces [13]. Wasserstein GAN formulations with Gradient Penalty (WGAN-GP) have been proposed to improve gradient behavior and convergence stability [14]. Moreover, attention mechanisms have demonstrated the ability to model long-range dependencies and improve structural coherence in generative tasks [15,16].

More recently, diffusion-based generative models have significantly advanced single-view 3D reconstruction. Approaches such as Zero-1-to-3 [17] and DreamFusion [18] leverage diffusion priors to synthesize novel views or optimize neural radiance fields, enabling high-fidelity implicit 3D representations. Similarly, Shap-E [19] and Point-E [20] demonstrate the effectiveness of diffusion-guided implicit and point-based generation frameworks. While these approaches achieve impressive visual realism, they typically rely on implicit representations or multi-stage optimization pipelines, which differ from voxel-based end-to-end adversarial reconstruction frameworks.

Beyond architectural components, effective optimization strategies play a critical role in volumetric generation. Adaptive learning rate scheduling techniques [21,22], ensemble learning approaches [23,24], and computational optimization methods such as automatic mixed precision have been shown to improve efficiency and training robustness in deep neural networks [25].

Despite these advances, several challenges persist in single-view voxel reconstruction: (i) limited geometric cues in RGB-only inputs, (ii) adversarial training instability in high-dimensional volumetric generation, and (iii) variance in reconstruction quality across training epochs. Addressing these issues requires a carefully stabilized generative framework that integrates complementary strategies within a unified reconstruction pipeline.

To this end, we propose SWAGE-3D (Spectral Wasserstein Attention Generative Ensemble for 3D Modeling), a stabilized RGB-D VAE-GAN framework designed for single-view voxel reconstruction. The proposed approach integrates depth-assisted feature encoding, Wasserstein adversarial training with spectral regularization, attention-based structural modeling, and checkpoint-based ensemble inference. Rather than

introducing entirely new building blocks, SWAGE-3D systematically integrates and validates complementary stabilization strategies to improve reconstruction consistency and volumetric coherence.

Contributions

The main contributions of this work are summarized as follows:

- Depth-Assisted RGB-D Encoding: A four-channel RGB-D input strategy is introduced by integrating monocular depth maps into the reconstruction pipeline. A ResNet18 encoder is adapted to accommodate depth information via mean-initialized channel expansion, enabling stable transfer learning.
- Stabilized Adversarial Volumetric Training: A Wasserstein GAN with Gradient Penalty is combined with spectral normalization to improve convergence stability and control Lipschitz continuity in high-dimensional voxel generation.
- Attention-Augmented Generator: A self-attention mechanism is incorporated to model long-range spatial dependencies, enhancing structural coherence in reconstructed voxel grids.
- Checkpoint-Based Ensemble Inference: An IoU-weighted ensemble strategy is employed to reduce prediction variance and improve inference robustness.
- Comprehensive Experimental Validation: Extensive ablation studies, sensitivity analysis, and multi-metric evaluations are conducted to quantify the individual and collective impact of the proposed components.

2 Related Work

Recent advancements in 3D reconstruction have been shaped by diverse representation paradigms and learning strategies, ranging from voxel-based modeling to implicit function learning and depth-assisted inference mechanisms. This section organizes prior studies according to representation paradigms and methodological focus.

2.1 Voxel-Based and Generative 3D Reconstruction

Voxel representations have long been adopted for structured 3D modeling due to their regular grid formulation and compatibility with convolutional neural networks. Early works explored visual similarity and rotation-invariant normalization for 3D retrieval [26,27].

The emergence of deep generative modeling significantly advanced volumetric synthesis. Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) established probabilistic and adversarial paradigms for high-dimensional data generation. Building upon these foundations, Wu et al. introduced 3D-GAN and 3D-VAE-GAN [28], enabling volumetric generation from both latent codes and 2D images. Recent advancements have sought to unify these paradigms; for instance, UniRecGen [29] integrates feed-forward reconstruction with diffusion-based generation to achieve highly consistent multi-view 3D models.

Recurrent aggregation mechanisms were later proposed for multi-view reconstruction in 3D-R2N2 [30]. Improvements in adversarial stability and scalable generative modeling were introduced through Wasserstein formulations like 3D-IWGAN [31] and geometry-aware 3D GANs [32]. Additional voxel-based generative and classification studies provided further architectural insights into 3D representation learning and recognition [33–37]. In particular, orientation-aware modeling strategies also enriched the design space of 3D deep learning frameworks [38].

More recent voxel-based frameworks focus on improved feature fusion and long-range dependency modeling. Pix2Vox [39] introduced context-aware fusion for combining coarse reconstructions, while TMVNet [40] leveraged Transformer-based encoders for enhanced multi-view feature aggregation. Similarly, VoxFormer [41] demonstrated the power of sparse voxel transformers for camera-based volumetric

scene completion. To further refine these architectures, SS3DNet-AF [42] proposed an attention-based fusion mechanism that enhances the network's focus on relevant geometric details from a single view. Similarly, recent efforts have directly upgraded multi-view pipelines by integrating multi-head attention refiners to reduce boundary prediction errors and capture intricate structural nuances [43]. Attention-enhanced voxel models such as SV3D-CDFF [44] and Semantic Voxel Structure (SVS) [45] further addressed discrepancies between image and voxel domains.

2.2 Alternative Representations: Point-Based and Implicit Methods

To overcome voxel resolution limitations, alternative 3D representations have been explored. Point-cloud-based generative models employed Earth Mover's Distance and deep autoencoding strategies for shape generation and completion [46,47]. DescriptorNet introduced an energy-based probabilistic model for 3D pattern synthesis [48].

Implicit representations further advanced continuous shape modeling. Occupancy Networks [49] represented 3D objects as continuous decision functions, enabling high-resolution reconstruction without explicit voxel discretization. IF-NET [50] extended implicit learning for shape completion. While these methods, alongside recent accelerated rendering advancements [51,52], alleviate discretization constraints, voxel representations remain advantageous for direct convolutional processing and adversarial volumetric learning.

Beyond implicit and point-based paradigms, studies have explored deformable surface modeling and graph-based generative approaches for structured 3D synthesis. Multi-chart surface parameterization and surface-oriented generation methods demonstrated that complex 3D geometry can be reconstructed through mesh- or chart-based learning strategies [53–55]. Studies on dynamic reconstruction, semantic shape modeling, and volumetric autoencoding supported the feasibility of learning structured geometric representations under diverse architectural assumptions [56–59]. While these approaches offer advantages in continuous surface modeling, they often require predefined topology constraints or specialized deformation priors. In contrast, voxel-based representations provide a regular volumetric grid structure that facilitates stable adversarial training and unified convolutional processing within a single-view reconstruction pipeline.

Recent advances in point-based and geometric learning have further expanded 3D representation capabilities. Transformer-based architectures and Vision Transformer (ViT) adaptations have been applied to point cloud processing and large-scale generalizable reconstruction to model global spatial dependencies more effectively [60,61]. In parallel, geometric neural operator frameworks have emerged as powerful tools for learning structured mappings in high-dimensional geometric domains [62,63]. These developments provide promising alternatives for geometric representation learning, complementing voxel-based approaches in scenarios where continuous or sparse representations are preferred. Transformer-based 3D reconstruction has also been extended toward neural implicit surface modeling. For example, SparseNeuS [64] integrates transformer architectures with neural implicit surfaces to enhance geometric consistency across sparse views, demonstrating improved structural coherence in continuous representations.

2.3 Depth Integration and Stabilization Strategies

Depth information has increasingly been incorporated to enhance geometric inference. Multi-view supervision methods such as Differentiable Ray Consistency (DRC) [65] demonstrated that volumetric predictions can be learned without explicit 3D labels. Recent advances in monocular depth estimation, particularly Depth Anything V2 [66], have significantly improved robustness and generalization across diverse environments.

Architectural refinements have also focused on stability and feature modeling. Self-attention mechanisms and Transformer-based encoders have been introduced to better capture long-range spatial dependencies in 3D reconstruction [67–69]. Ensemble learning strategies have been widely adopted to reduce prediction variance and enhance robustness in deep neural networks.

Despite substantial progress, several challenges persist in single-view voxel reconstruction, including limited geometric cues in RGB-only inputs, adversarial instability in high-dimensional volumetric generation, and reconstruction variance across training epochs. Although prior studies have addressed these issues individually, their combined and systematically stabilized integration within a unified RGB-D voxel-based generative framework has received comparatively limited investigation. In this context, the present study emphasizes the controlled integration and empirical validation of complementary stabilization strategies for depth-assisted voxel reconstruction.

3 Spectral Wasserstein Attention Generative Ensemble for 3D Model (SWAGE-3D)

3.1 Overview of SWAGE-3D

Unlike our previous VAE-based approach [70], which primarily focused on transfer learning for latent representation stabilization, the proposed SWAGE-3D framework systematically integrates adversarial Wasserstein training, spectral normalization, self-attention modules, and ensemble learning to address stability and geometric coherence in single-view voxel reconstruction. SWAGE-3D framework builds upon the foundational work of Wu et al. [28] on 3D-VAE-GAN, introducing several key innovations to enhance both training stability and the quality of generated 3D models. The overall architecture, illustrated in Fig. 2, consists of three primary network components that operate in a unified voxel-based generative pipeline.

The architecture integrates three primary network components that work together to deliver superior performance. The Encoder Network serves as the initial processing unit, taking 2D input images that include both RGB and depth information and mapping them into a latent space characterized by mean (μ) and variance (σ) parameters. By incorporating depth maps alongside RGB data, the encoder captures richer geometric details, enabling more accurate and informative latent representations. The Generator Network then transforms latent vectors sampled from this space into detailed 3D voxel models using transposed 3D convolutions and a novel self-attention mechanism. This mechanism enhances the model's ability to capture long-range structural coherence. Finally, the Discriminator Network evaluates the generated 3D models to differentiate between real scanned objects and synthetic samples. By providing precise adversarial feedback based on the Wasserstein distance, this component guides the generator in producing outputs that are increasingly indistinguishable from real-world data. The algorithm for the proposed model is shown in Algorithm 1.

3.2 RGB-D Encoder

The SWAGE-3D framework employs a depth-assisted RGB-D encoding strategy to enrich geometric feature representation from single-view inputs. Instead of relying solely on RGB images, monocular depth maps are generated and fused as an additional channel, forming a four-channel input representation. Depth maps are extracted from the input RGB images using the Depth Anything V2 model, which provides robust monocular depth estimation across diverse scenes. Each RGB image of size $224 \times 224 \times 3$ is augmented with its corresponding depth map to construct a $224 \times 224 \times 4$ RGB-D tensor. This integration enables the encoder to capture complementary geometric cues that are not explicitly encoded in color information alone.

All inputs are normalized using ImageNet statistics for the RGB channels, while the depth channel is normalized using its dataset-wide mean and standard deviation to maintain scale consistency. A ResNet18

backbone is adopted as the encoder component. Since the standard ResNet18 architecture is pre-trained on 3-channel RGB images, its first convolutional layer is modified to accept 4-channel RGB-D inputs. To preserve transfer learning benefits, the pre-trained weights corresponding to the three RGB channels are retained. The weights of the additional depth channel are initialized as the mean of the RGB channel weights: $W_{\text{depth}} = \frac{1}{3}(W_R + W_G + W_B)$. This initialization allows the depth channel to start from a balanced representation without disrupting learned feature distributions. The modified encoder outputs mean μ and log-variance $\log \sigma^2$ parameters, from which latent vectors are sampled via the reparameterization trick:

$$z = \mu_\phi(x) + \sigma_\phi(x) \odot \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I) \tag{1}$$

where ε stands for the random noise drawn from the standard normal distribution. This stochastic encoding enables variational learning while preserving stable gradient propagation during training.

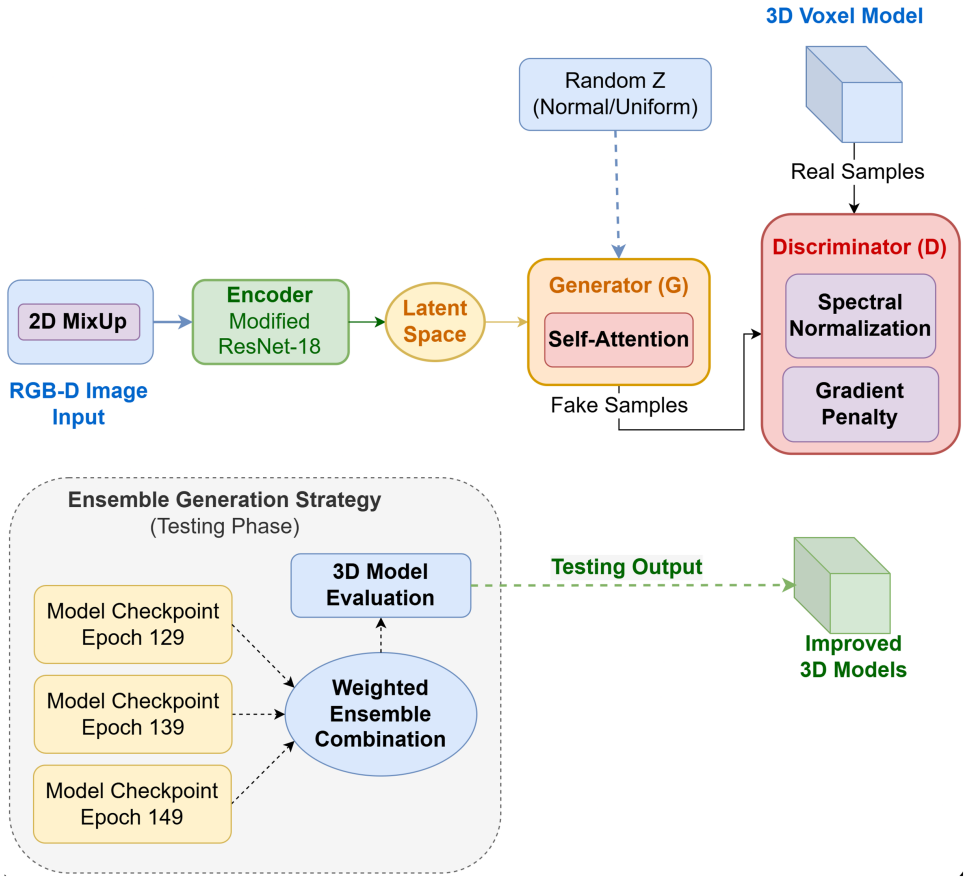

Figure 2: Overall architecture of the proposed SWAGE-3D framework.

3.3 Generative Backbone and Adversarial Stabilization

The generative backbone consists of a 3D convolutional generator and a Wasserstein-based discriminator. The generator transforms the sampled latent vector into a $32 \times 32 \times 32$ voxel grid. To address instability commonly observed in standard GAN training, a Wasserstein GAN with Gradient Penalty (WGAN-GP) formulation is adopted [71]. The discriminator estimates the Wasserstein distance between real and generated voxel distributions. Its objective is defined as:

$$\mathcal{L}_D = \mathbb{E}_{\hat{x}\sim P_g}[D(\hat{x})] - \mathbb{E}_{x\sim P_r}[D(x)] + \lambda_{GP}\,\mathbb{E}_{\tilde{x}\sim P_{\tilde{x}}}\left(\|\nabla_{\tilde{x}} D(\tilde{x})\|_2 - 1\right)^2 \tag{2}$$

Although the Wasserstein distance formulation based on the Kantorovich–Rubinstein (KR) duality provides a theoretically sound objective, its practical implementation remains an approximation [72]. In theory, the KR dual formulation considers the supremum over all 1-Lipschitz functions. However, the critic is restricted to a parameterized neural network with finite capacity. The learned critic provides only an approximation of the true Wasserstein-1 distance rather than its exact computation [72].

To mitigate these limitations, we adopt a hybrid stabilization strategy combining spectral normalization (global Lipschitz control) with a reduced-weight gradient penalty term (local gradient regularization). This dual constraint balances critic expressiveness and stability, improving approximation smoothness without severely restricting model capacity. This approach serves as a stabilized practical realization of Wasserstein-based adversarial learning. In our generator network, a self-attention mechanism is incorporated after the third layer to model long-range spatial dependencies, ensuring global consistency in the generated volumes.

Algorithm 1: Proposed SWAGE-3D model: spectral wasserstein attention generative ensemble

Input: Training dataset $\mathcal{D}$, batch size m, critic iterations $n_{critic} = 5$, GP weight λ_{GP}
Output: Ensemble of Generator models
if *use_SN is True* **then**
  Apply spectral normalization to D;
  $\lambda_{GP} \leftarrow 1.0$;
for $epoch = 1$ **to** *n_epochs* **do**
  for *batch (images, depths, voxels) in* $DataLoader(\mathcal{D})$ **do**
    /* Encoder-Generator forward */
    $z\_mu, z\_var \leftarrow E(concat(images, depths))$;
    $z \leftarrow reparameterize(z\_mu, z\_var)$;
    $fake\_voxels \leftarrow G(z)$;
    /* With self-attention */
    /* Discriminator update */
    for $t = 1$ **to** n_{critic} **do**
      /* WGAN-GP loss */
      $d\_loss \leftarrow D(fake\_voxels).mean() - D(voxels).mean()$;
      $d\_loss \leftarrow d\_loss + compute\_gradient\_penalty(D, voxels, fake\_voxels, \lambda_{GP})$;
      Update D parameters;
    /* Generator-Encoder update */
    $g\_loss \leftarrow -D(fake\_voxels).mean() + \lambda_{recon} \times MSE(fake\_voxels, voxels)$;
    $e\_loss \leftarrow \lambda_{recon} \times MSE(fake\_voxels, voxels) + \lambda_{kl} \times KL(z\_mu, z\_var)$;
    Update G and E parameters;
  if *epoch in key checkpoints* **then**
    Save checkpoint of networks;
/* Ensemble creation from specific checkpoints */
Load models from epochs $\{129, 139, 149\}$;
for *each model* G_i *in ensemble* **do**
  $iou_i \leftarrow calculate\_iou(G_i(z\_test), voxels\_test)$;
$ensemble\_output \leftarrow weighted\_combination(G_{129}, G_{139}, G_{149}, weights = iou\_based)$;
return $ensemble\_output$

3.4 Self-Attention Module

To enhance structural coherence in volumetric generation, a self-attention mechanism is incorporated into the generator. While convolutional layers effectively capture local spatial patterns, their receptive fields remain limited. In voxel-based reconstruction tasks, long-range spatial dependencies are critical for maintaining global object consistency. The self-attention module enables direct interactions between distant spatial regions within the volumetric feature maps. For an intermediate 3D feature tensor $F \in \mathbb{R}^{C\times H\times W\times D}$, query, key, and value representations are obtained via $1 \times 1 \times 1$ convolutions. The attention map is computed based on query-key similarity, and the resulting representation is combined with the original feature tensor through a residual connection: $F' = \gamma \cdot \text{Attention}(Q, K, V) + F$. Here, γ is a learnable scalar parameter initialized to zero to stabilize early training.

3.5 Ensemble Strategy

To enhance inference robustness and mitigate performance fluctuations across training epochs, a checkpoint-based ensemble strategy is employed during the testing phase. Instead of relying on a single trained model, multiple late-stage checkpoints are aggregated to produce a more stable and accurate voxel reconstruction. Model performance was monitored using the IoU metric, and a stabilization plateau was observed between epochs 125 and 149, indicating convergence with minor oscillations. From this stable region, epochs 129, 139, and 149 were selected to construct the ensemble.

Let $G_i(x)$ denote the voxel prediction produced by the i-th checkpoint model, and let w_i represent its validation IoU score. The final ensemble prediction is computed as a weighted aggregation: $\hat{V}(x) = \frac{\sum w_i G_i(x)}{\sum w_i}$. Following aggregation, a voxel occupancy threshold $\tau = 0.5$ is applied to obtain the final binary volumetric reconstruction. This ensemble mechanism reduces variance introduced by stochastic training dynamics and improves overall reconstruction accuracy.

3.6 Composite Loss Function and Training Dynamics

The optimization of SWAGE-3D is formulated as a multi-stage composite objective that jointly regulates latent space regularity, volumetric reconstruction fidelity, and adversarial distribution alignment. The framework simultaneously optimizes three interacting networks: the encoder (E), the generator (G), and the discriminator (D), using a coordinated update strategy that bridges Variational Autoencoder (VAE) and Generative Adversarial Network (GAN) paradigms.

The primary objective of the VAE stream is to ensure that the encoder-generator pair can effectively map a single-view RGB-D input to its corresponding 3D geometry. To ensure voxel-wise consistency between predicted grids and ground-truth occupancy, a Mean Squared Error (MSE) reconstruction loss is employed:

$$\mathcal{L}_{rec} = \frac{1}{N}\sum_{i=1}^{N}(V_i^{pred} - V_i^{gt})^2 \tag{3}$$

where V^{pred} and V^{gt} denote the predicted and ground-truth voxel grids, respectively. Simultaneously, the encoder is regularized using the Kullback–Leibler (KL) divergence to enforce the latent distribution to follow a unit Gaussian prior $p(z) \sim \mathcal{N}(0, I)$ [73]:

$$\mathcal{L}_{KL} = -\frac{1}{2}\sum(1 + \log\sigma^2 - \mu^2 - \sigma^2) \tag{4}$$

This dual objective allows the encoder to capture a robust, continuous representation of the 3D shapes while the generator learns the fundamental volumetric mapping.

To enhance the realism and structural sharpness of the generated voxels beyond simple point-wise matching, an adversarial stream is integrated. Under the Wasserstein formulation, the discriminator acts as a critic that estimates the Earth-Mover distance between the generated distribution P_g and the real data distribution P_r. Its objective is defined as:

$$\mathcal{L}_D = \mathbb{E}_{\hat{x}\sim P_g}[D(\hat{x})] - \mathbb{E}_{x\sim P_r}[D(x)] + \lambda_{GP}\mathcal{L}_{GP} \tag{5}$$

where $\mathcal{L}_{GP}$ represents the gradient penalty term enforcing the 1-Lipschitz continuity constraint [74]. In our implementation, spectral normalization is applied to the critic's layers to provide global Lipschitz control, allowing the gradient penalty coefficient λ_{GP} to be reduced to 1.0. This prevents the vanishing gradient problem and stabilizes the training of the generator.

The technical novelty of the SWAGE-3D training procedure lies in the coordinated update of the generator. The generator is not merely trained to fool the discriminator but is explicitly forced to satisfy both the VAE reconstruction constraint and the GAN adversarial requirement. The total objectives for the encoder ($\mathcal{L}_E$) and generator ($\mathcal{L}_G$) are defined as:

$$\mathcal{L}_E = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{KL}\mathcal{L}_{KL} \tag{6}$$

$$\mathcal{L}_G = \lambda_{rec}\mathcal{L}_{rec} - \mathbb{E}_{\hat{x}\sim P_g}[D(\hat{x})] \tag{7}$$

By minimizing $\mathcal{L}_G$, the generator learns to produce voxels that are geometrically accurate relative to the input image (via $\mathcal{L}_{rec}$) while simultaneously conforming to the global distribution of real 3D objects (via the adversarial term). In all experiments, the weights are set to $\lambda_{rec} = 1.0$ and $\lambda_{KL} = 1.0$. To maintain equilibrium in this complex adversarial landscape, the discriminator is updated $n_{critic} = 5$ times for every single update of the encoder and generator, ensuring a reliable approximation of the Wasserstein distance before each generative refinement step.

4 Experiments

4.1 Experimental Setup

4.1.1 Dataset

This study utilizes the ShapeNet dataset, a widely recognized and comprehensive repository of 3D models across various object categories. ShapeNet contains a total of 55 different object classes, making it one of the most extensively used datasets in the literature for 3D object generation and reconstruction tasks [75]. In alignment with prior studies, this research focuses on a subset of 13 objects selected from the ShapeNet dataset. These objects were chosen to ensure compatibility with existing benchmarks and facilitate meaningful comparisons.

The ShapeNet dataset provides 3D model files for each object, which can be voxelized using the Binvox method. The voxel term refers to a volumetric pixel, representing a point in the 3D environment where a cube is either present (state 1) or absent (state 0). Using tools developed by the ShapeNet team and other researchers, 2D images of the models can be generated from multiple viewpoints as desired. Fig. 3 demonstrates an example of a 3D voxel representation alongside its corresponding 2D image for each object used in this study. The dataset is divided into 80% for training and 20% for testing for each object, ensuring a balanced evaluation framework.

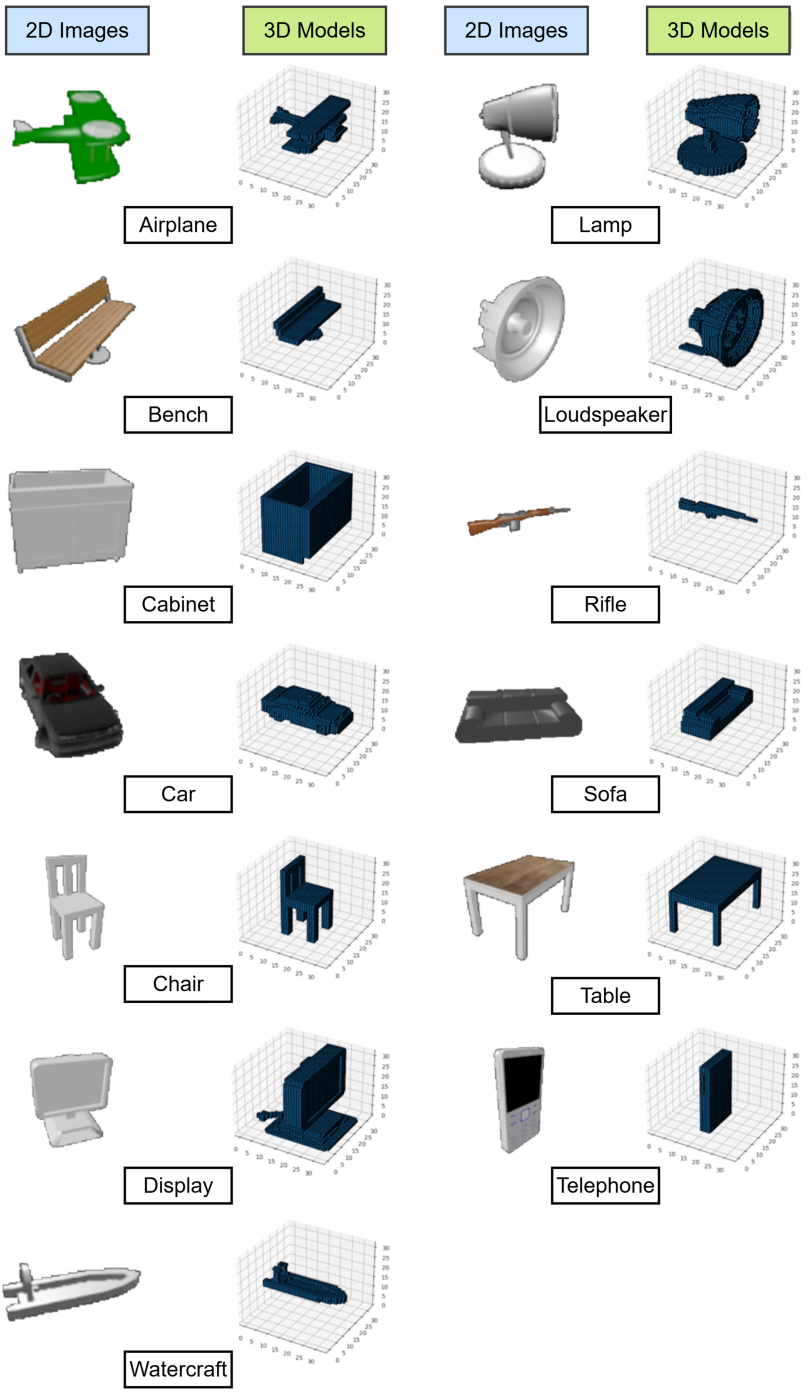

Figure 3: 2D images and corresponding 3D models of 13 objects selected from the ShapeNet dataset.

To ensure compatibility with the ResNet18 architecture, 2D images were resized to dimensions of 224 × 224 × 3. Here, the first two dimensions (224 × 224) represent the spatial resolution (horizontal and vertical pixel count), while the third dimension corresponds to the RGB channels (Red, Green, Blue). Each channel is represented by 8 bits, resulting in a total of 24 bits per image. While some datasets include an additional alpha channel (RGBA) to represent transparency, this study focuses exclusively on RGB images.

For 3D models, a voxelized representation is employed. Each model is expressed within a 32 × 32 × 32 grid (resolution), where cubes are either added or omitted at specific points in the 3D space. This binary representation ensures efficient processing while preserving the structural integrity of the objects.

Following preprocessing procedures, 2D input images with dimensions of 224 × 224 × 3 and their corresponding 3D voxel objects of size 32 × 32 × 32 were prepared. Depth maps were extracted from the 2D images using the Depth Anything V2 architecture [66] and integrated as a fourth channel into the input. These depth maps are represented as 8-bit grayscale images, expanding the input dimensions to 224 × 224 × 4.

The incorporation of depth maps was motivated by several potential improvements. Specifically, this enhancement aims to provide the architecture with additional geometric information, spatial depth cues, and improved object boundary delineation. Furthermore, it mitigates challenges related to illumination variations, enriches feature representation, and strengthens the discriminative capabilities of the model. Fig. 4 illustrates example inputs along with their corresponding depth maps extracted using the Depth Anything V2 architecture.

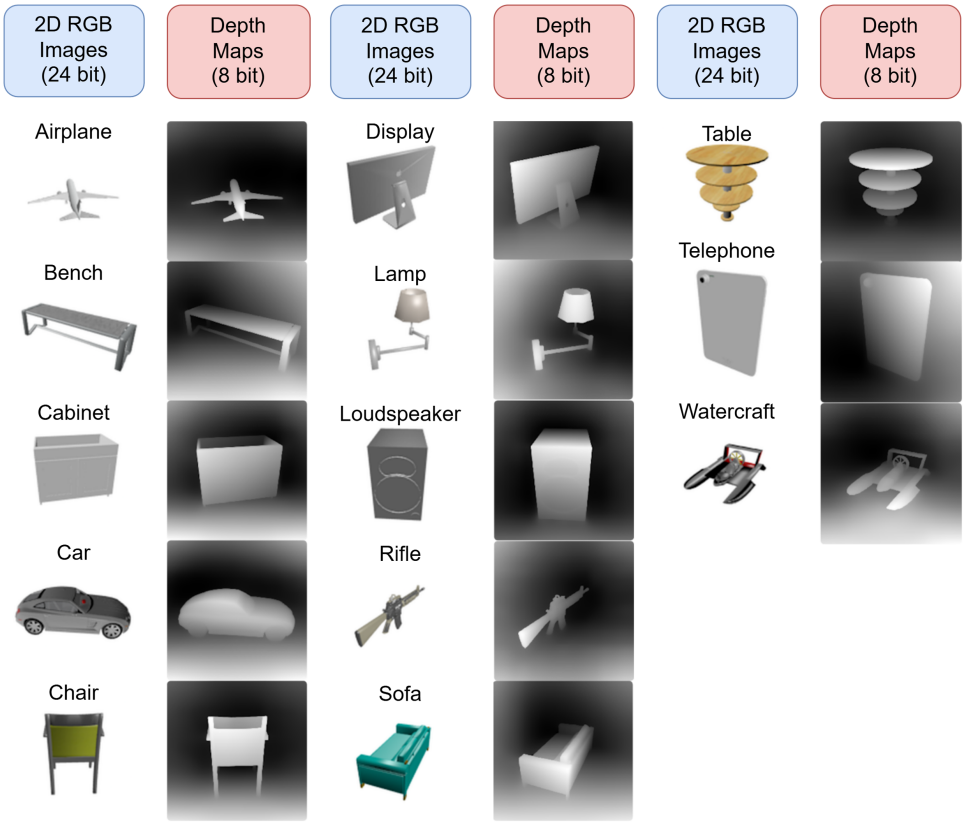

Figure 4: 2D images and corresponding depth maps using Depth Anything V2.

During the integration of depth maps, specific normalization parameters were employed. The mean (0.330) and standard deviation (0.236) values used for depth channel normalization were empirically calculated across the entire training subset of the ShapeNet dataset to ensure optimal distribution for the ResNet18 encoder.

All experiments were conducted on the ShapeNetCore dataset using the 13 object categories described in this section. The dataset was divided into 80% training and 20% testing samples for each category to ensure

balanced evaluation across object types. The 3D models were voxelized at a resolution of $32 \times 32 \times 32$, while the corresponding 2D input images were resized to 224×224. Monocular depth maps were generated for each RGB image and concatenated as a fourth channel, forming an RGB-D input representation.

4.1.2 Implementation Details

The SWAGE-3D framework was implemented using the PyTorch deep learning library. All experiments were conducted on a workstation equipped with an Intel i7-4790 CPU, 16 GB RAM, and a single NVIDIA GTX 980 Ti GPU.

To ensure reproducibility and training stability, we adopted a specific weight initialization and data augmentation strategy. The ResNet18 encoder backbone, pre-trained on ImageNet, was modified to accept four-channel RGB-D inputs. The weights for the additional depth channel were initialized as the mean of the pre-trained RGB weights, $W_{\text{depth}} = \text{avg}(W_R, W_G, W_B)$, to preserve the benefits of transfer learning. During training, 2D MixUp data augmentation was applied to the input images with an interpolation probability of $p = 0.5$ and a Beta distribution parameter of $\alpha = 0.2$ to mitigate mode collapse.

The final hyperparameters, determined through empirical validation and sensitivity analysis, are summarized in Table 1. All models were trained for 150 epochs with a batch size of 32. The latent space dimensionality was strictly set to $d_z = 200$ to balance representational capacity and training stability.

Table 1: Final training configuration and hyperparameters.

Category	Parameter	Value
Optimizer (Adam)	Generator Learning Rate (η_G)	2.5×10^{-3}
	Discriminator Learning Rate (η_D)	1.0×10^{-3}
	Encoder Learning Rate (η_E)	1.0×10^{-4}
	Betas (β_1, β_2)	$(0.5, 0.9)$
Loss Weights	Reconstruction Weight (λ_{rec})	1.0
	KL Divergence Weight (λ_{KL})	1.0
	Gradient Penalty Weight (λ_{GP})	1.0*
	Critic Iterations (n_{critic})	5
Stability	Latent Dimension (d_z)	200
	Classifier Threshold (d_{thresh})	0.8
	MixUp Prob/Alpha (p/α)	0.5/0.2
	Occupancy Threshold (τ)	0.5

Note: *Reduced from 10.0 to 1.0 when Spectral Normalization is enabled.

Optimization was performed using the Adam optimizer. To prevent the discriminator from overpowering the generator, a balancing threshold was implemented: the discriminator was only updated if its classification accuracy was below $d_{\text{thresh}} = 0.8$. Following the WGAN-GP strategy, the critic was updated five times per generator update ($n_{\text{critic}} = 5$).

The composite loss function integrates voxel reconstruction loss (MSE), KL divergence ($\lambda_{\text{KL}} = 1.0$), and adversarial feedback. In the baseline WGAN-GP configuration, λ_{GP} was set to 10, but was reduced to 1.0 when spectral normalization was enabled to prevent over-constraining the critic. Model performance was monitored via Intersection over Union (IoU). As shown in Fig. 5, stabilization occurs after epoch 125, motivating the selection of late-stage checkpoints (epochs 129, 139, and 149) for ensemble construction.

For quantitative evaluation, probabilistic voxel outputs were binarized using a fixed occupancy threshold of $\tau = 0.5$.

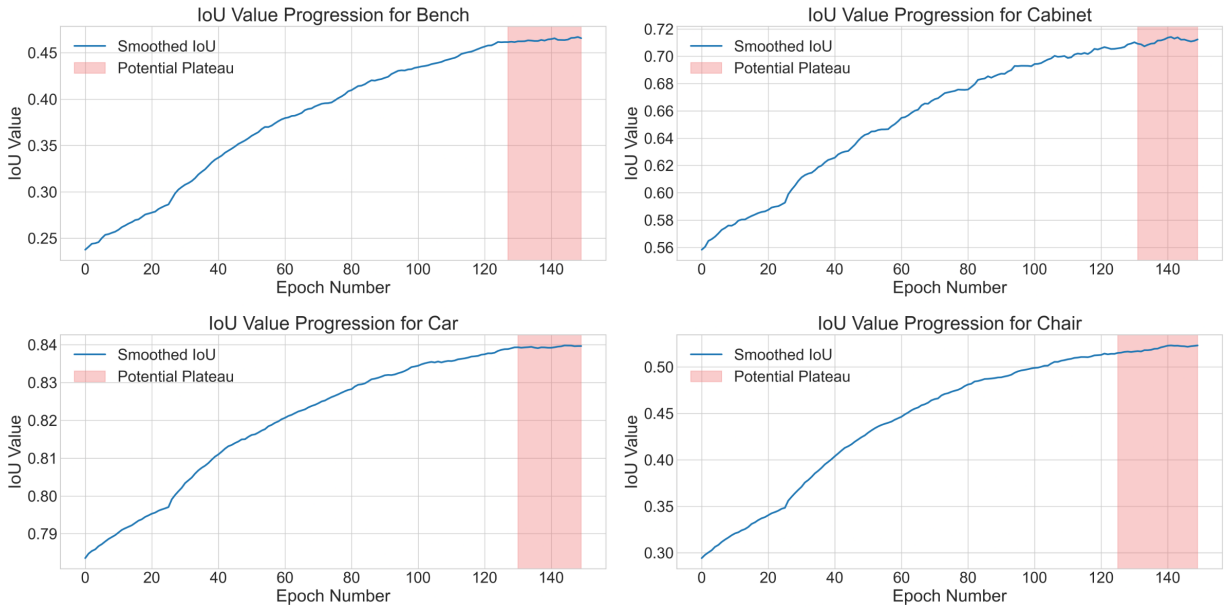

Figure 5: Results and potential plateau regions for 150 epochs of training IoU for the objects Bench, Cabinet, Car, and Chair.

4.2 Sensitivity Analysis of Loss Weights

To evaluate the robustness of the SWAGE-3D framework, we conducted a sensitivity study on the primary loss weights, λ_{rec} and λ_{KL}, using the Loudspeaker category as a representative sample. We tested scaling factors of ±2× relative to the default configuration. Due to computational constraints, each configuration was trained for 30 epochs to observe early convergence trends and training stability. The results are summarized in Table 2.

Table 2: Sensitivity analysis of loss weights on Loudspeaker category (evaluated at epoch 30).

Configuration	λ_{rec}	λ_{KL}	Mean IoU	Stability
Default (Optimal)	1.0	1.0	0.6676	**Stable**
Low Reconstruction	0.5	1.0	0.6617	Stable
High Reconstruction	2.0	1.0	**0.6766**	Minor Oscillations
Low KL Regularization	1.0	0.5	0.6600	High Variance
High KL Regularization	1.0	2.0	0.6695	Stable (Blurred)

Note: Bold values indicate the best performance.

The results demonstrate that SWAGE-3D exhibits remarkable stability across different objective weightings, with IoU scores remaining within a narrow margin (<2%). While the 'High Reconstruction' ($\lambda_{rec} = 2.0$) configuration yielded a slightly higher IoU of 0.6766 at this stage, it introduced minor oscillations in the discriminator loss, which could potentially destabilize the adversarial balance during full-scale 150-epoch training. Conversely, reducing λ_{KL} to 0.5 led to increased variance in the Wasserstein critic's feedback, confirming that adequate latent regularization is essential for smooth adversarial mapping. Doubling the KL

weight (λ_{KL} = 2.0) maintained stability but resulted in slightly smoother, less detailed voxel grids (posterior collapse). Based on these observations, the λ_{rec} = 1.0 and λ_{KL} = 1.0 configuration was selected as the optimal balance for long-term training, ensuring both high geometric fidelity and consistent adversarial equilibrium across all object categories.

4.3 Evaluation Metrics

To quantitatively evaluate the reconstruction performance of SWAGE-3D and the compared methods, multiple complementary metrics were employed to assess volumetric overlap, structural accuracy, and geometric distance consistency. The primary evaluation metric is the Intersection over Union (IoU), which measures the volumetric overlap between predicted and ground-truth voxel grids. IoU is defined as:

$$IoU = \frac{\sum_i (V_i^{pred} \cdot V_i^{gt})}{\sum_i (V_i^{pred} + V_i^{gt} - V_i^{pred} \cdot V_i^{gt})} \tag{8}$$

where V^{pred} and V^{gt} denote the predicted and ground-truth voxel occupancy grids, respectively. This metric provides a rigorous assessment of spatial consistency between reconstructed and reference 3D shapes. For all evaluations, a voxel occupancy threshold of τ = 0.5 was applied to binarize the generator outputs.

While IoU evaluates total volumetric overlap, it may not fully capture the precision of reconstructed surface details. Therefore, the F-score (F_1) is additionally reported to provide a balanced measure of reconstruction fidelity by calculating the harmonic mean of precision and recall based on voxel occupancy. The F-score is computed as:

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{9}$$

In this context, precision represents the accuracy of the predicted occupied voxels, while recall indicates the fraction of the ground-truth structure successfully recovered by the model. This metric is particularly effective in evaluating the reconstruction quality of thin and complex structures where volumetric overlap alone might be insufficient to reflect structural integrity.

To further quantify the geometric consistency between reconstructed and ground-truth shapes, the Chamfer Distance (*CD*) is employed. Given two point sets *P* and *G* derived from the coordinates of occupied voxels in the predicted and ground-truth grids, respectively, *CD* is defined as:

$$CD(P, G) = \frac{1}{|P|} \sum_{p \in P} \min_{g \in G} \|p - g\|_2^2 + \frac{1}{|G|} \sum_{g \in G} \min_{p \in P} \|g - p\|_2^2 \tag{10}$$

Chamfer Distance evaluates bidirectional nearest-neighbor consistency and captures geometric deviations that influence surface accuracy even when they do not significantly affect the overall voxel overlap. By jointly reporting IoU, F-score, and Chamfer Distance, a comprehensive assessment of volumetric reconstruction accuracy, structural precision, and geometric consistency is provided.

4.4 Quantitative Comparison with Existing Methods

To evaluate its performance, SWAGE-3D is compared with several representative 3D reconstruction methods, including 3D-R2N2 [30], which employs Recurrent Neural Networks (RNNs) for single- or multi-view reconstruction; 3D-VAE-GAN [28], a hybrid VAE–GAN framework for volumetric generation;

Differentiable Ray Consistency (DRC) [65], which leverages multi-view supervision; Pix2Mesh [76], a mesh-based generative model; and Occupancy Networks (ONet) [49], which represent 3D shapes as continuous implicit functions.

In addition, we include more recent voxel-based approaches such as TMVNet [40], which utilizes a Transformer-based 3D encoder for modeling long-range volumetric dependencies, and Pix2Vox-A [39], which adopts a context-aware multi-view voxel refinement strategy.

Table 3 provides a detailed comparison of SWAGE-3D with these methods based on category-specific and average Intersection over Union (IoU) scores evaluated on the ShapeNet dataset. Although TMVNet and Pix2Vox-A achieve higher overall IoU values on ShapeNet, it is important to contextualize these results within architectural and training paradigm differences. Both TMVNet and Pix2Vox-A are designed around multi-view supervision and feature fusion mechanisms, enabling stronger geometric consistency during training. In contrast, SWAGE-3D operates under a single-view generative framework enhanced with depth integration and adversarial learning.

Table 3: Comparison of different 3D model generation methods based on IoU metric.

Category	Method							
	3D-R2N2 [30]	3DVAEGAN [28]	DRC [65]	Pix2Mesh [76]	ONet [49]	TMVNet [40]	Pix2Vox-A [39]	Ours –
Airplane	0.513	0.420	0.571	0.420	0.571	0.691	0.684	0.572
Bench	0.421	0.340	0.453	0.323	0.485	0.659	0.616	0.401
Cabinet	0.716	0.600	0.635	0.664	0.733	0.853	0.792	0.726
Car	0.798	0.760	0.755	0.552	0.737	0.870	0.854	0.835
Chair	0.466	0.360	0.469	0.396	0.501	0.721	0.567	0.460
Display	0.468	0.400	0.419	0.490	0.471	0.595	0.537	0.419
Lamp	0.381	0.320	0.415	0.323	0.371	0.534	0.443	0.423
Loudspeaker	0.662	0.590	0.609	0.599	0.647	0.712	0.714	0.693
Rifle	0.544	0.540	0.608	0.402	0.474	0.783	0.615	0.463
Sofa	0.628	0.570	0.606	0.613	0.680	0.701	0.709	0.662
Table	0.513	0.330	0.424	0.395	0.506	0.660	0.601	0.524
Telephone	0.661	0.680	0.413	0.661	0.720	0.801	0.776	0.731
Watercraft	0.513	0.480	0.556	0.397	0.530	0.685	0.594	0.544
Average	0.560	0.491	0.533	0.479	0.571	0.712	0.661	0.573

TMVNet leverages a Transformer-based 3D encoder to model long-range volumetric dependencies across multiple views, which provides stronger structural priors during reconstruction. Similarly, Pix2Vox-A employs a context-aware refinement strategy that explicitly aggregates multi-view voxel predictions, leading to improved geometric completeness.

By comparison, SWAGE-3D prioritizes generative modeling stability and distribution learning through a hybrid VAE-WGAN framework. While adversarial learning enhances structural realism and diversity, it does not directly optimize IoU in a purely supervised regression sense. Therefore, the slightly lower IoU values should be interpreted as a trade-off between generative robustness and direct voxel regression performance rather than a deficiency in reconstruction capability.

The results reveal a competitive landscape where different methods excel in specific categories. Across categories, SWAGE-3D consistently improves over the baseline 3D-VAE-GAN and achieves its most pronounced gains on rigid object classes (e.g., Car and Telephone), suggesting that the proposed stabilization and depth fusion improve reconstruction reliability. For instance, among single-view baselines, SWAGE-3D demonstrates a substantial margin in the Car category, achieving an IoU score of 0.835, and remains highly competitive even against multi-view methods. This exceptional performance in rigid, man-made categories is largely driven by the depth-aware encoder, which successfully resolves the depth ambiguity of flat surfaces that plagues purely RGB-based moels. While multi-view architectures like TMVNet establish the upper bound in categories such as Bench, Cabinet, and Rifle, SWAGE-3D achieves the highest average IoU among the compared single-view generative voxel-based baselines. This underscores the framework's ability to deliver high-quality reconstructions across a wide range of object geometries, offering a robust balance between detail and structural coherence.

It is worth noting that while implicit representation methods like ONet excel in organic shapes due to their continuous nature, SWAGE-3D demonstrates superior performance in rigid, geometric objects (e.g., Cars with 0.835 IoU vs. ONet's 0.737). This indicates that our voxel-based approach, fortified with depth guidance and attention, is particularly effective for preserving the structural integrity of man-made objects with defined planar surfaces.

While the proposed SWAGE-3D model demonstrates superior performance in volumetric objects such as cars and loudspeakers, a slight performance drop is observed in categories characterized by thin and complex structures, such as Rifle and Bench. This limitation is attributed to the fixed voxel resolution (32^3) and the regularization effects of spectral normalization, which may occasionally treat fine structural details as high-frequency noise during the reconstruction process. Increasing the voxel resolution in future works could effectively mitigate this limitation.

To further validate the effectiveness of SWAGE-3D, we compare its performance with the baseline 3DVAEGAN model using the IoU metric, as illustrated in Fig. 6. When evaluating individual objects, SWAGE-3D demonstrates significant improvements. Specifically, it achieves a 15.2% higher IoU for the Airplane category and a 19.4% improvement for the Table category compared to 3DVAEGAN. For objects like Display and Telephone, SWAGE-3D shows modest gains of 1.9% and 5.1%, respectively. However, in the Rifle category, 3DVAEGAN outperforms SWAGE-3D by 7.7%. Compared to the baseline 3D-VAE-GAN evaluated under the same single-view voxel reconstruction protocol, SWAGE-3D improves the average IoU from 0.491 to 0.573, corresponding to a 16.7% relative gain.

Fig. 7 provides qualitative insights into the performance of SWAGE-3D. The figure illustrates the input RGB image, the associated depth map (colorized for enhanced visibility), the voxelized model generated by SWAGE-3D, the corresponding ground truth voxelized model from the dataset, and the intersection volume between the predicted and ground truth voxels. To highlight performance variations, examples representing the highest (Car), average (Airplane), and lowest (Bench) IoU values achieved by SWAGE-3D are included. These qualitative results corroborate the quantitative findings, demonstrating the framework's ability to produce highly accurate and structurally coherent 3D models.

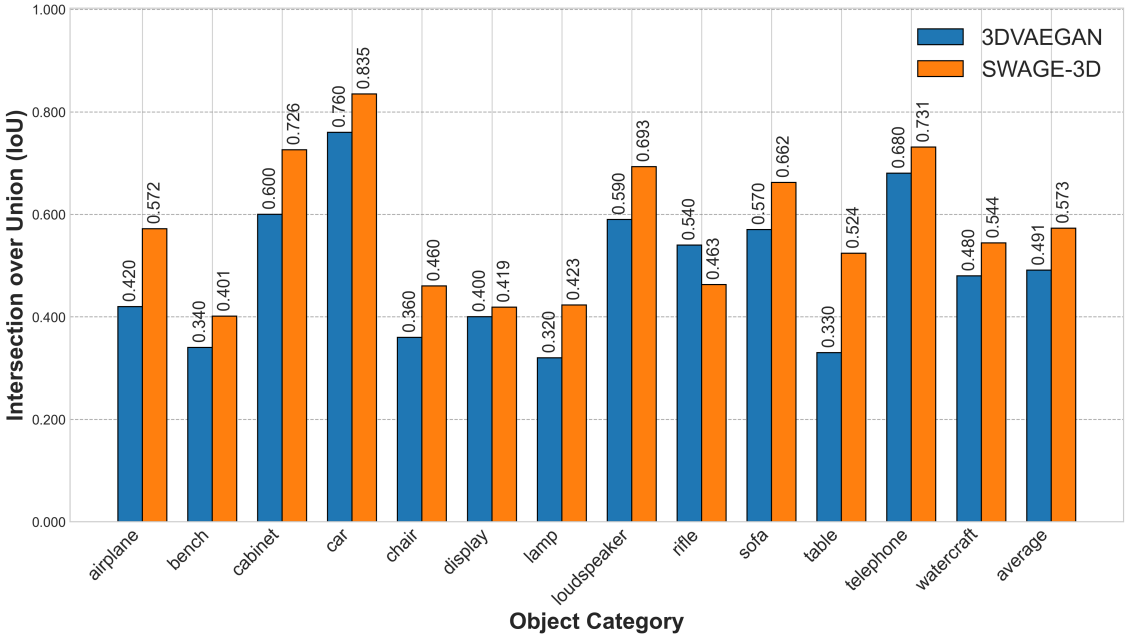

Figure 6: Comparison of 3DVAEGAN and SWAGE-3D models on IoU metric for 13 objects.

Fig. 8 provides a qualitative and quantitative comparison of 3D-VAE-GAN, 3D-R2N2, Pix2Mesh, and SWAGE-3D. Depth maps are included in the inputs section but are only utilized for SWAGE-3D. Ground Truth refers to the actual 3D voxelized models in the dataset. Analysis of the 3D results highlights the superior performance of SWAGE-3D. Similar to the qualitative results presented earlier, SWAGE-3D demonstrates strong performance for Airplane, Car, Lamp, Loudspeaker, Table, and Telephone. For the Airplane, it produces a model very similar to the real one, though minor extraneous voxels appear outside the structure. For the Bench, while there are deficiencies in the railing parts, the basic skeleton is correctly constructed.

For the Cabinet, despite some issues with the arm sections, the fundamental structure is accurately produced. For the Car, the generated model closely resembles the real object. For the Chair, errors exist in the backrest and seating areas, but the overall production is successful. For the Display, the model includes the stand and is well-produced. The Lamp is reconstructed with high success. For the Loudspeaker, despite minor errors in the front section, the output is satisfactory. For the Rifle, intermediate voxels are missing. This limitation is attributed to the fixed voxel resolution (32^3) and the aggressive regularization of Spectral Normalization, which occasionally treats very thin structures as high-frequency noise. However, the model still correctly localizes the main body, suggesting that higher resolutions in future work could resolve this. The Sofa is produced with great success, with only minor voxel deficiencies in the seating area. For the Table, the generated model closely matches the real one, though minor errors persist in the foot sections. The Telephone is reconstructed with 100% accuracy. For the Watercraft, the model performs well, despite some deficiencies in the front and upper sections. Overall, SWAGE-3D achieves excellent results compared to other models in the literature, particularly excelling in specific object categories.

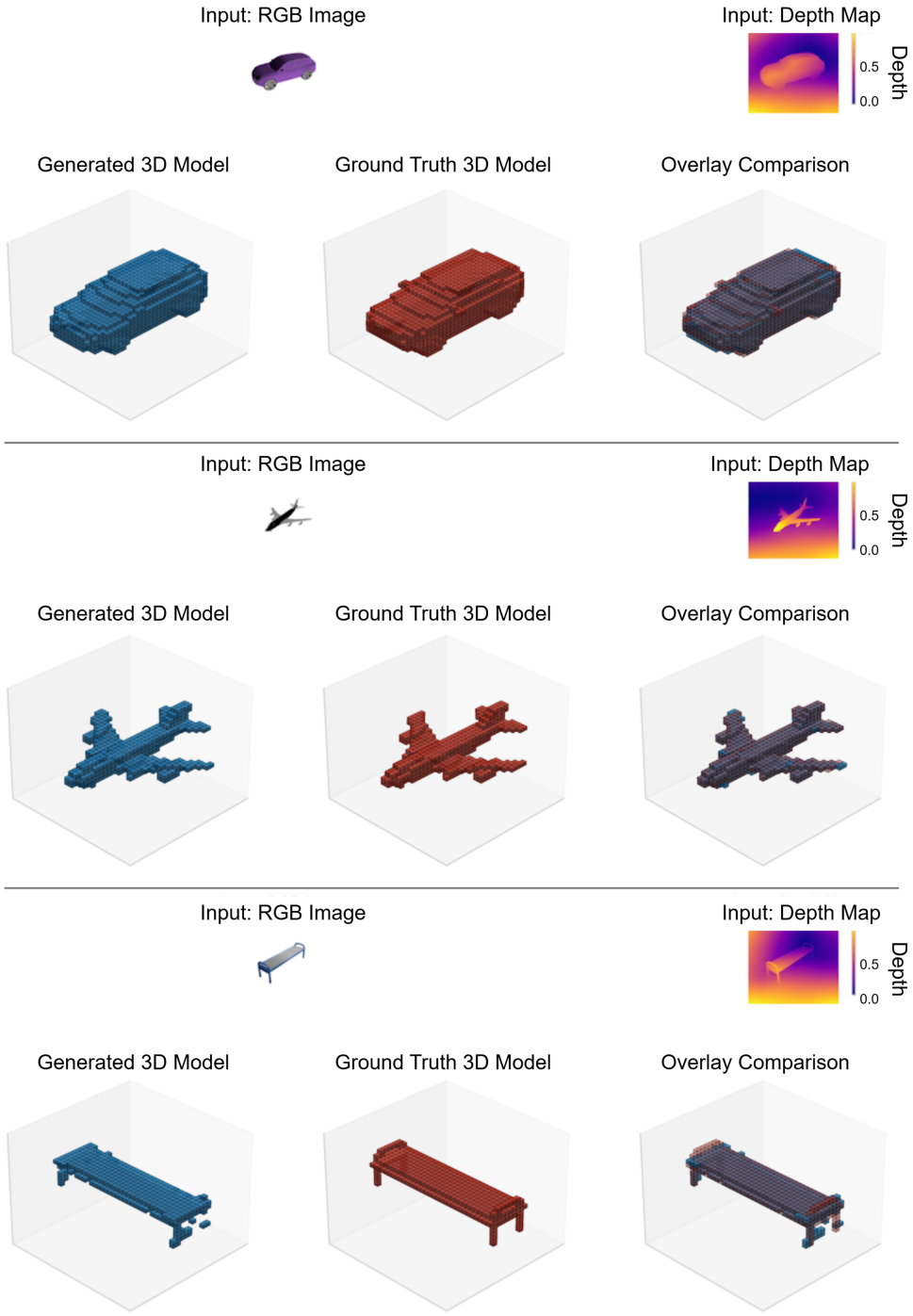

Figure 7: Qualitative performance of SWAGE-3D: input RGB and depth maps (colorized), predicted voxels, ground truth voxels, and their intersection for high (Car), average (Airplane), and low (Bench) IoU results.

4.5 Incremental Ablation Study

To systematically quantify the individual contributions of the integrated techniques within the SWAGE-3D framework, an incremental ablation study was conducted. Due to the high computational demand of full-scale training, this study was performed on six representative categories (Airplane, Bench, Car, Lamp,

Loudspeaker, and Telephone) that exhibit diverse geometric characteristics. To ensure a fair and rigorous evaluation, all ablation variants were trained under a separate, strictly unified experimental setup to isolate the relative gain of each component. While absolute values may vary slightly from the peak performance results reported in Table 3 due to the stochastic nature of adversarial training, the performance trends across configurations remain entirely consistent.

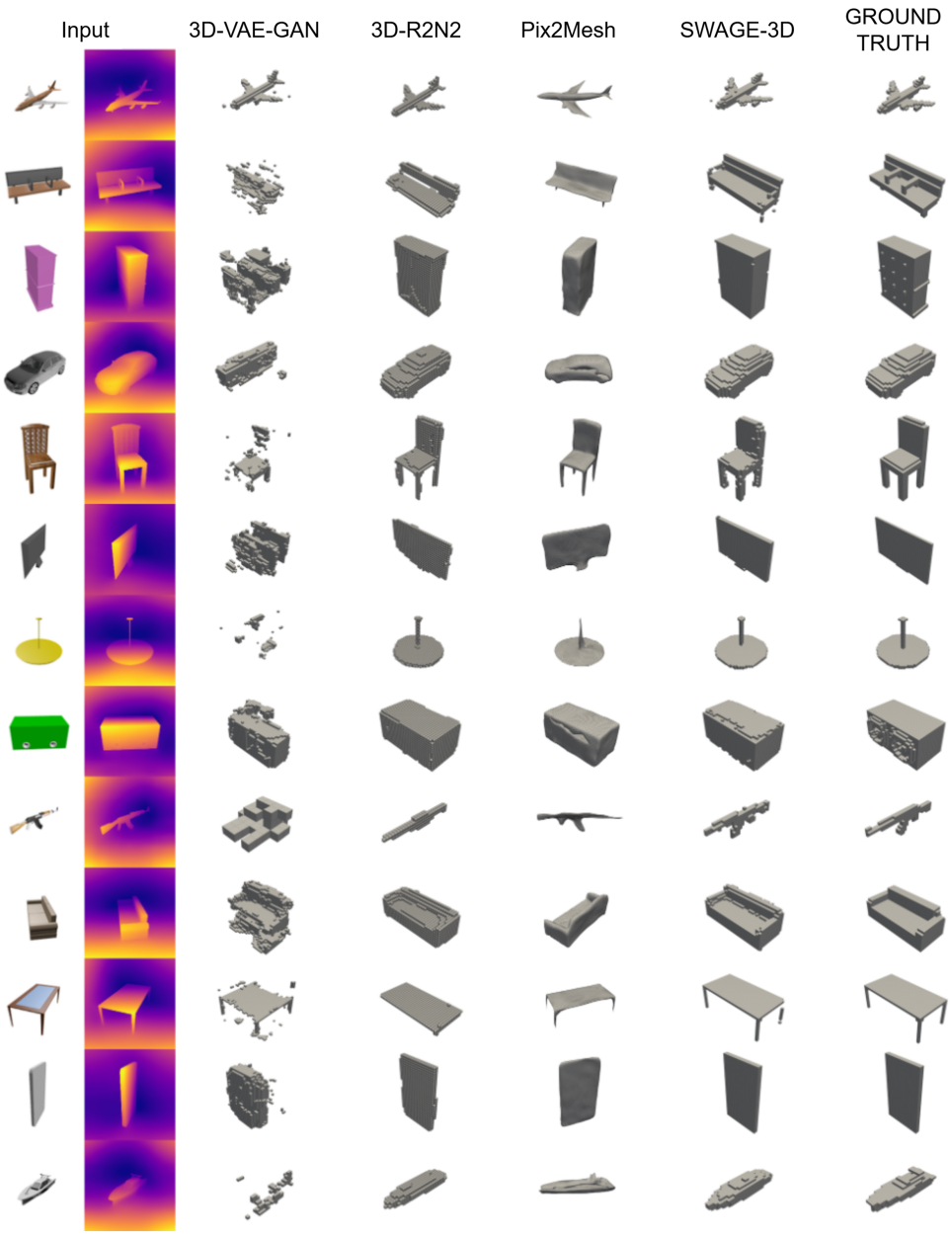

Figure 8: Comparison of 3D-VAE-GAN, 3D-R2N2, Pix2Mesh, and SWAGE-3D across object categories.

Five model variants were evaluated: (1) a baseline 3D-VAE-GAN using 3-channel RGB input, (2) the integration of monocular depth maps (RGB-D), (3) the addition of Wasserstein GAN loss with Gradient Penalty (WGAN-GP), (4) the inclusion of Self-Attention modules, and (5) the final weighted

ensemble inference. The results across Mean IoU, F-Score, and Chamfer Distance (CD) are summarized in Tables 4–6, respectively.

Table 4: Incremental ablation study results for Mean IoU across six representative categories.

Configuration	Airplane	Bench	Car	Lamp	Loudspeaker	Telephone	Mean
(1) Baseline (RGB)	0.557	0.380	0.828	0.340	0.677	0.729	0.585
(2) + Depth Map	0.570	0.385	0.825	**0.431**	0.689	**0.730**	**0.605**
(3) + WGAN-GP	**0.581**	0.387	**0.829**	0.395	0.689	0.727	0.601
(4) + Attention	0.568	**0.392**	0.812	0.410	**0.691**	0.717	0.598
(5) + Ensemble	0.567	**0.392**	0.816	0.412	0.688	0.715	0.598

Note: Bold values indicate the best performance.

Table 5: Incremental ablation study results for F-Score across six representative categories.

Configuration	Airplane	Bench	Car	Lamp	Loudspeaker	Telephone	Mean
(1) Baseline (RGB)	0.715	0.550	**0.906**	0.506	0.807	**0.843**	0.721
(2) + Depth Map	0.725	0.554	0.904	**0.598**	0.815	**0.843**	**0.740**
(3) + WGAN-GP	**0.734**	0.555	**0.906**	0.563	0.816	0.841	0.736
(4) + Attention	0.724	**0.562**	0.896	0.580	**0.817**	0.834	0.735
(5) + Ensemble	0.723	0.561	0.899	0.579	0.815	0.833	0.735

Note: Bold values indicate the best performance.

Table 6: Incremental ablation study results for Chamfer Distance (CD).

Configuration	Airplane	Bench	Car	Lamp	Speaker	Telephone	Mean
(1) Baseline (RGB)	0.955	1.618	**0.214**	10.015	0.925	0.423	2.358
(2) + Depth Map	0.885	1.597	0.219	4.157	0.874	**0.407**	1.357
(3) + WGAN-GP	0.874	1.825	0.219	5.841	0.841	**0.399**	1.667
(4) + Attention	0.845	**1.469**	0.233	4.252	**0.839**	0.413	1.342
(5) + Ensemble	**0.826**	1.499	0.227	**4.051**	0.872	0.423	**1.316**

Note: Bold values indicate the best performance.

The systematic ablation reveals that the integration of depth maps (Config 2) is the primary driver for recovering geometric fidelity, resolving ambiguities where RGB data alone is insufficient. This is most evident in the Lamp category, where depth maps facilitated a substantial reduction in total geometric error. Subsequent architectural refinements via WGAN-GP and Self-Attention (Configs 3 and 4) focused on enhancing global shape consistency and boundary sharpness. While point-wise metrics like Mean IoU exhibit minor fluctuations, a common characteristic of GAN-based models where distributional realism is prioritized over local MSE, the structural metrics (F-Score and CD) indicate a consistent downward trend in geometric error across configurations. This study confirms that the cumulative success of SWAGE-3D arises from the synergistic combination of depth-assisted geometric initialization, stabilized adversarial learning, and robust ensemble inference.

4.6 Resolution Analysis (32^3 and 64^3)

To address the scalability of SWAGE-3D and evaluate its performance at higher voxel densities, we conducted a comparative analysis between the standard 32^3 resolution and a higher 64^3 grid size. This experiment was specifically performed on the Bench category, as it contains thin and complex structures that provide a rigorous test for high-resolution 3D generation. To ensure a fair assessment of computational efficiency and scaling behavior, both models were trained for 50 epochs under consistent experimental conditions.

The results, summarized in Table 7, demonstrate that SWAGE-3D effectively scales to higher resolutions while maintaining structural integrity. While the 64^3 resolution significantly increases the learning complexity and the state space of the voxel grid, the model successfully captures the essential geometric properties of the target shapes. As expected, the computational demand increases with resolution; specifically, GPU memory usage rose from 0.41 to 1.34 GB, and the training time per epoch increased from 1.08 to 11.25 min. Although the IoU and F-score values for the 64^3 resolution are slightly lower than those of 32^3 within the fixed 50-epoch training budget the model's ability to handle eight times the voxel density confirms its architectural robustness and scalability for high-fidelity 3D reconstruction.

Table 7: Resolution analysis of SWAGE-3D under different voxel grid sizes. Performance is evaluated on the ShapeNet test split (Bench category) after 50 training epochs. Computational cost is measured on a single NVIDIA GTX 980 Ti GPU.

Resolution	IoU	F-Score (↑)	Chamfer Dist. (↓)	GPU Memory (GB)	Training Time (min)
32^3	0.382	0.552	1.5416	0.41	1.08
64^3	0.343	0.506	3.0343	1.34	11.25

4.7 Discussion

The experimental findings provide a comprehensive assessment of the proposed SWAGE-3D framework across reconstruction accuracy, geometric consistency, stabilization behavior, and scalability. The results collectively indicate that the integration of depth-assisted encoding and stabilized adversarial training contributes to improved reconstruction robustness under a single-view voxel-based setting.

The quantitative comparison on ShapeNet demonstrates that SWAGE-3D consistently outperforms the baseline 3D-VAE-GAN across multiple object categories in terms of IoU. The most pronounced improvements are observed in rigid and structurally well-defined categories such as Car and Telephone, suggesting that depth-guided encoding effectively mitigates depth ambiguity in planar and symmetric structures. This indicates that the RGB-D fusion strategy plays a central role in improving geometric reliability when only a single image is available.

The incremental ablation study further clarifies the relative contributions of architectural components. The depth map integration yields the most substantial improvement in mean IoU and F-score, confirming that geometric cues extracted from monocular depth estimation significantly enhance volumetric inference. The inclusion of WGAN-GP and spectral normalization primarily contributes to training stabilization and distribution alignment rather than dramatic IoU gains. While the improvement in volumetric overlap metrics is moderate, the reduction in Chamfer Distance suggests improved geometric smoothness and surface coherence. Similarly, the self-attention module provides marginal gains in voxel overlap but contributes to enhanced structural consistency, particularly reflected in distance-based evaluation. The ensemble

strategy offers limited incremental improvement in IoU but helps reduce variance and improve robustness during inference.

Taken together, these findings suggest that performance gains do not stem from a single dominant architectural modification alone, but rather from a controlled combination of complementary stabilization and geometric enhancement strategies. The improvements are more pronounced in structural reliability and geometric consistency than in raw volumetric overlap alone.

The inclusion of F-score and Chamfer Distance provides a more nuanced understanding of reconstruction behavior. While IoU captures volumetric overlap, F-score reflects precision-recall balance of occupancy prediction, and Chamfer Distance quantifies geometric proximity. The observed trend, where some configurations yield modest IoU changes but measurable Chamfer improvements, indicates that adversarial stabilization and attention mechanisms primarily refine surface quality rather than drastically altering voxel occupancy ratios. This highlights the importance of multi-metric evaluation when analyzing generative volumetric models.

The resolution analysis comparing 32^3 and 64^3 voxel grids reveals a critical trade-off between geometric granularity and computational feasibility. Although the 64^3 configuration increases representational capacity, it leads to reduced IoU and F-score under the current hardware constraints, while significantly increasing memory consumption and training time. The cubic growth in voxel space dramatically enlarges the optimization landscape, making stable adversarial training more challenging on a single GTX 980 Ti GPU. These findings indicate that while the architecture is technically scalable to higher resolutions, effective optimization at 64^3 resolution requires either stronger regularization strategies or more advanced computational resources. Therefore, 32^3 remains a practical and balanced resolution choice for stable single-view adversarial voxel reconstruction under moderate hardware settings.

Despite its improvements, SWAGE-3D has several limitations. First, the voxel-based representation inherently suffers from discretization artifacts and cubic computational complexity. Although adversarial learning improves structural realism, thin structures and fine details remain difficult to reconstruct at low resolutions. While higher resolutions are theoretically feasible, they introduce substantial computational overhead and optimization instability. Second, the model relies on monocular depth estimation as a preprocessing step. Errors in predicted depth maps directly propagate to the latent representation, potentially limiting reconstruction fidelity in visually ambiguous or textureless regions.

Third, while the incremental ablation study demonstrates the contribution of each component, the magnitude of improvement beyond depth integration is moderate. This suggests that geometric cues play a more dominant role than adversarial refinement in single-view voxel reconstruction under the current configuration. Finally, SWAGE-3D is evaluated under a single-view generative paradigm. Multi-view transformer-based methods achieve higher absolute IoU values by leveraging stronger geometric supervision. The proposed framework prioritizes stabilization and distribution-aware volumetric generation rather than purely supervised regression optimization.

Future research may explore hybrid voxel-implicit representations to alleviate discretization constraints, adaptive resolution strategies to dynamically allocate voxel density, and more advanced regularization techniques to stabilize higher-resolution adversarial training. Additionally, integrating uncertainty-aware depth estimation or jointly optimizing depth and voxel generation in an end-to-end manner may further enhance geometric consistency.

5 Conclusion

This study was motivated by persistent challenges in 3D object generation, particularly the limitations of existing 3D-VAE-GAN architectures in capturing fine-grained detail, ensuring training stability, and using spatial information from single-view inputs. To overcome these obstacles, we proposed SWAGE-3D, a novel framework that integrates several state-of-the-art enhancements including depth map integration, MixUp data augmentation, a ResNet18-based encoder, spectral normalization, self-attention mechanisms, Wasserstein GAN with gradient penalty (WGAN-GP), and ensemble model testing. These contributions collectively aim to address gaps in the literature by combining complementary techniques that enhance both the stability and fidelity of voxel-based 3D reconstructions. Our results confirm that while individual components like Attention or Depth Integration are powerful on their own, their combined application creates a necessary stability for voxel-based GANs, yielding a 16.7% improvement without requiring complex multi-stage training pipelines.

In this study, we introduced SWAGE-3D, an advanced framework designed to overcome the limitations of existing 3D generative architectures by integrating depth-enhanced input, self-attention mechanisms, ensemble learning, and training stabilization techniques such as spectral normalization and WGAN-GP. Unlike prior studies that focused primarily on the standard 3D-VAE-GAN, our work systematically incorporates and improves upon a broader range of recent state-of-the-art (SOTA) models, offering a more robust and accurate solution for single-view 3D object reconstruction.

Extensive evaluations were conducted on the ShapeNet dataset using the Intersection over Union (IoU) metric. Compared to the baseline 3D-VAE-GAN under identical single-view voxel reconstruction settings, SWAGE-3D achieves a relative IoU improvement of 16.7%, increasing the average IoU from 0.491 to 0.573. Beyond this, while multi-view architectures achieve higher overall bounds, SWAGE-3D demonstrated superior or comparable performance when benchmarked against leading single-view baselines such as ONet (0.571), 3D-R2N2 (0.560), and DRC (0.533). These results highlight SWAGE-3D's capacity to produce structurally coherent, detail-rich 3D reconstructions.

This performance gain can be attributed to several architectural and training innovations. The integration of Depth Anything V2-based depth maps as a fourth input channel enhanced the geometric reasoning of the encoder. The ResNet18-based encoder, adapted for 4-channel input, enabled efficient and accurate feature extraction. The attention-augmented generator improved global context understanding, while the WGAN-GP + spectral normalization combination addressed training instability, leading to smoother convergence. Moreover, ensemble testing across selected model checkpoints further improved inference robustness and generalization.

Looking ahead, future research directions include extending SWAGE-3D to handle real-world noisy or occluded inputs, and exploring alternative representations such as point clouds or implicit surfaces to improve scalability and detail fidelity. Additionally, incorporating multi-view fusion, cross-modal learning, and transformer-based encoders may further elevate reconstruction accuracy and enable broader applicability across domains like autonomous navigation, AR/VR, and medical imaging.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Zafer Serin: Conceptualization, Methodology, Software, Validation, Visualization, Writing—Original Draft; Cihan Karakuzu: Conceptualization, Supervision, Writing—Review & Editing; Uğur Yüzgeç: Conceptualization, Methodology, Supervision, Writing—Review & Editing. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The experiments in this study were conducted using the ShapeNetCore dataset, publicly available at the official Stanford repository (https://shapenet.org/). Access to the dataset requires registration through the official website. In this work, we utilized a subset of 13 object categories following the common experimental protocol adopted in prior voxel-based reconstruction studies. The source code, pre-trained models, and detailed instructions to reproduce the results will be made publicly available on GitHub upon acceptance of the manuscript.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ding Y, Yang H, Xu C, Zhang C, Li L. DGGR-Net: single-image 3D reconstruction from complex backgrounds via graph-based refinement and difference-guided fusion. J King Saud Univ Comput Inf Sci. 2025;37:222. doi:10.1007/s44443-025-00251-8.
2. Kazhdan M, Hoppe H. Screened poisson surface reconstruction. ACM Trans Graph. 2013;32(3):29. doi:10.1145/2487228.2487237.
3. Rusu RB, Cousins S. 3D is here: point cloud library (PCL). In: 2011 IEEE International Conference on Robotics and Automation (ICRA). Piscataway, NJ, USA: IEEE; 2011. p. 1–4. doi:10.1109/icra.2011.5980567.
4. Botsch M, Kobbelt L. A remeshing approach to multiresolution modeling. In: Proceedings of the 2004 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing. New York, NY, USA: ACM; 2004. p. 185–92. doi:10.1145/1057432.1057457.
5. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86(11):2278–324. doi:10.1109/5.726791.
6. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NeurIPS). Red Hook, NY, USA: Curran Associates, Inc.; 2012. p. 1097–105.
7. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2016. p. 770–8. doi:10.1109/cvpr.2016.90.
8. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. 2015.
9. Sun C, Shrivastava A, Singh S, Gupta A. Revisiting unreasonable effectiveness of data in deep learning era. In: 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE; 2017. p. 843–52. doi:10.1109/iccv.2017.97.
10. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. Mixup: beyond empirical risk minimization. arXiv:1710.09412. 2018.
11. Kingma DP, Welling M. Auto-encoding variational Bayes. arXiv:1312.6114. 2014.
12. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Advances in Neural Information Processing Systems (NeurIPS). Red Hook, NY, USA: Curran Associates, Inc.; 2014. p. 2672–80.
13. Chakraborty T, Reddy KSU, Naik SM, Panja M, Manvitha B. Ten years of generative adversarial nets (GANs): a survey of the state-of-the-art. Mach Learn Sci Technol. 2024;5(1):011001. doi:10.1088/2632-2153/ad1f77.
14. Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning. Brookline, MA, USA: PMLR; 2017. p. 214–23.
15. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS). Red Hook, NY, USA: Curran Associates, Inc.; 2017. p. 5998–6008.
16. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proceeding of the International Conference on Learning Representations (ICLR); 2015 May 7–9; San Diego, CA, USA.

17. Liu R, Wu R, Van Hoorick B, Tokmakov P, Zakharov S, Vondrick C. Zero-1-to-3: zero-shot one image to 3D object. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE; 2023. p. 9264–75. doi:10.1109/iccv51070.2023.00853.
18. Poole B, Jain A, Barron JT, Mildenhall B. Dreamfusion: text-to-3D using 2D diffusion. arXiv:2209.14988. 2022.
19. Jun H, Nichol A. Shap-E: generating conditional 3D implicit functions. arXiv:2305.02463. 2023.
20. Nichol A, Jun H, Dhariwal P, Mishkin P, Chen M. Point-E: a system for generating 3D point clouds from complex prompts. arXiv:2212.08751. 2022.
21. Smith LN. Cyclical learning rates for training neural networks. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway, NJ, USA: IEEE; 2017. p. 464–72. doi:10.1109/wacv.2017.58.
22. Loshchilov I, Hutter F. SGDR: stochastic gradient descent with warm restarts. In: ICLR 2017 (5th International Conference on Learning Representations); 2017 Apr 24–26; Toulon, France.
23. Dietterich TG. Ensemble methods in machine learning. In: Proceedings of the First International Workshop on Multiple Classifier Systems. Cham, Switzerland: Springer; 2000. p. 1–15. doi:10.1007/3-540-45014-9_1.
24. Sagi O, Rokach L. Ensemble learning: a survey. Wiley Interdiscip Rev Data Min Knowl Discov. 2018;8(4):e1249. doi:10.1002/widm.1249.
25. Striuk O, Kondratenko Y. Optimization strategy for generative adversarial networks design. Int J Comput. 2023;22(3):292–301. doi:10.47839/ijc.22.3.3223.
26. Chen DY, Tian XP, Shen YT, Ouhyoung M. On visual similarity based 3D model retrieval. Comput Graph Forum. 2003;22(3):223–32. doi:10.1111/1467-8659.00669.
27. Funkhouser T, Min P, Kazhdan M, Chen J, Halderman A, Dobkin D, et al. A search engine for 3D models. ACM Trans Graph. 2003;22(1):83–105. doi:10.1145/588272.588279.
28. Wu J, Zhang C, Xue T, Freeman B, Tenenbaum J. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In: Advances in Neural Information Processing Systems (NeurIPS). Red Hook, NY, USA: Curran Associates, Inc.; 2016. p. 82–90.
29. Huang Z, Chen J, Lin C, Hu C, Huang H, Yu Z, et al. UniRecGen: unifying multi-view 3D reconstruction and generation. arXiv:2604.01479. 2026.
30. Choy CB, Xu D, Gwak J, Chen K, Savarese S. 3D-R2N2: a unified approach for single and multi-view 3D object reconstruction. In: The European Conference on Computer Vision (ECCV). Cham, Switzerland: Springer; 2016. p. 628–44. doi:10.1007/978-3-319-46484-8_38.
31. Smith EJ, Meger D. Improved adversarial systems for 3D object generation and reconstruction. In: Proceeding of the Conference on Robot Learning (CoRL). Brookline, MA, USA: PMLR; 2017. p. 87–96.
32. Chan ER, Lin CZ, Chan MA, Nagano K, Pan B, De Mello S, et al. Efficient geometry-aware 3D generative adversarial networks. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2022. p. 16102–12. doi:10.1109/cvpr52688.2022.01565.
33. Maturana D, Scherer S. VoxNet: a 3D convolutional neural network for real-time object recognition. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway, NJ, USA: IEEE; 2015. p. 922–8. doi:10.1109/iros.2015.7353481.
34. Su H, Maji S, Kalogerakis E, Learned-Miller E. Multi-view convolutional neural networks for 3D shape recognition. In: 2015 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE; 2015. p. 945–53. doi:10.1109/iccv.2015.114.
35. Johns E, Leutenegger S, Davison AJ. Pairwise decomposition of image sequences for active multi-view recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2016. p. 3813–22. doi:10.1109/cvpr.2016.414.
36. Hegde V, Zadeh R. FusionNet: 3D object classification using multiple data representations. arXiv:1607.05695. 2016.
37. Brock A, Lim T, Ritchie JM, Weston N. Generative and discriminative voxel modeling with convolutional neural networks. arXiv:1608.04236. 2016.
38. Sedaghat N, Zolfaghari M, Amiri E, Brox T. Orientation-boosted voxel nets for 3D object recognition. arXiv:1604.03351. 2016.

39. Xie H, Yao H, Sun X, Zhou S, Zhang S. Pix2Vox: context-aware 3D reconstruction from single and multi-view images. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE; 2019. p. 2690–8. doi:10.1109/iccv.2019.00278.
40. Peng K, Islam R, Quarles J, Desai K. TMVNet: using transformers for multi-view voxel-based 3D reconstruction. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway, NJ, USA: IEEE; 2022. p. 221–9. doi:10.1109/cvprw56347.2022.00036.
41. Li Y, Yu Z, Choy C, Xiao C, Alvarez JM, Fidler S, et al. VoxFormer: sparse voxel transformer for camera-based 3D semantic scene completion. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2023. p. 9087–98. doi:10.1109/CVPR52729.2023.00877.
42. Shoukat MA, Sargano AB, Malyshev A, You L, Habib Z. SS3DNet-AF: a single-stage, single-view 3D reconstruction network with attention-based fusion. Appl Sci. 2024;14(23):11424. doi:10.3390/app142311424.
43. Lee K, Cho I, Yang B, Park U. Multi-head attention refiner for multi-view 3D reconstruction. J Imaging. 2024;10(11):268. doi:10.3390/jimaging10110268.
44. Xiong W, Huang F, Zhang H, Jiang M. 3D voxel reconstruction from single-view image based on cross-domain feature fusion. Expert Syst Appl. 2024;256(1):124957. doi:10.1016/j.eswa.2024.124957.
45. Li Y, Liu Z, Benes B, Zhang X, Guo J. SVDTree: semantic voxel diffusion for single image tree reconstruction. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2024. p. 4692–702. doi:10.1109/cvpr52733.2024.00449.
46. Fan H, Su H, Guibas LJ. A point set generation network for 3D object reconstruction from a single image. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2017. p. 2463–71. doi:10.1109/cvpr.2017.264.
47. Achlioptas P, Diamanti O, Mitliagkas I, Guibas L. Learning representations and generative models for 3D point clouds. In: The International Conference on Machine Learning (ICML). Brookline, MA, USA: PMLR; 2018. p. 40–9.
48. Xie J, Zheng Z, Gao R, Wang W, Zhu SC, Wu YN. Learning descriptor networks for 3D shape synthesis and analysis. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2018. p. 8629–38. doi:10.1109/cvpr.2018.00900.
49. Mescheder L, Oechsle M, Niemeyer M, Nowozin S, Geiger A. Occupancy networks: learning 3D reconstruction in function space. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2019. p. 4455–65. doi:10.1109/cvpr.2019.00459.
50. Chibane J, Alldieck T, Pons-Moll G. Implicit functions in feature space for 3D shape reconstruction and completion. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2020. p. 6968–79. doi:10.1109/cvpr42600.2020.00700.
51. Müller T, Evans A, Schied C, Keller A. Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans Graph. 2022;41(4):102. doi:10.1145/3528223.3530127.
52. Kerbl B, Kopanas G, Leimkühler T, Drettakis G. 3D Gaussian splatting for real-time radiance field rendering. ACM Trans Graph. 2023;42(4):139. doi:10.1145/3592433.
53. Ben-Hamu H, Maron H, Kezurer I, Avineri G, Lipman Y. Multi-chart generative surface modeling. ACM Trans Graph. 2018;37(6):215. doi:10.1145/3272127.3275052.
54. Groueix T, Fisher M, Kim VG, Russell BC, Aubry M. A papier-mâché approach to learning 3D surface generation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2018. p. 216–24. doi:10.1109/cvpr.2018.00030.
55. Litany O, Bronstein A, Bronstein M, Makadia A. Deformable shape completion with graph convolutional autoencoders. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2018. p. 1886–95. doi:10.1109/cvpr.2018.00202.
56. Bogo F, Romero J, Pons-Moll G, Black MJ. Dynamic FAUST: registering human bodies in motion. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2017. p. 5573–82. doi:10.1109/cvpr.2017.591.

57. Yang Y, Yu Y, Zhou Y, Du S, Davis J, Yang R. Semantic parametric reshaping of human body models. In: 2014 2nd International Conference on 3D Vision. Piscataway, NJ, USA: IEEE; 2014. p. 41–8. doi:10.1109/3dv.2014.47.
58. Girdhar R, Fouhey DF, Rodriguez M, Gupta A. Learning a predictable and generative vector representation for objects. In: The European Conference on Computer Vision (ECCV). Cham, Switzerland: Springer; 2016. p. 484–99. doi:10.1007/978-3-319-46466-4_29.
59. Sharma A, Grau O, Fritz M. Vconv-dae: deep volumetric shape learning without object labels. In: The European Conference on Computer Vision (ECCV). Cham, Switzerland: Springer; 2016. p. 236–50. doi:10.1007/978-3-319-49409-8_20.
60. Zhao H, Jiang L, Jia J, Torr PHS, Koltun V. Point transformer. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE; 2021. p. 16239–48. doi:10.1109/iccv48922.2021.01595.
61. Hong Y, Zhang K, Gu J, Bi S, Zhou Y, Liu D, et al. LRM: large reconstruction model for single image to 3D. arXiv:2311.04400. 2024.
62. Li Z, Kovachki N, Azizzadenesheli K, Liu B, Bhattacharya K, Stuart A, et al. Fourier neural operator for parametric partial differential equations. arXiv:2010.08895. 2021.
63. Kovachki N, Li Z, Liu B, Azizzadenesheli K, Bhattacharya K, Stuart A, et al. Neural operator: learning maps between function spaces with applications to pdes. J Mach Learn Res. 2023;24(89):1–97.
64. Long X, Lin C, Wang P, Komura T, Wang W. Sparseneus: fast generalizable neural surface reconstruction from sparse views. In: The European Conference on Computer Vision (ECCV). Cham, Switzerland: Springer; 2022. p. 210–27. doi:10.1007/978-3-031-19824-3_13.
65. Tulsiani S, Zhou T, Efros AA, Malik J. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2017. p. 209–17. doi:10.1109/cvpr.2017.30.
66. Yang L, Kang B, Huang Z, Zhao Z, Xu X, Feng J, et al. Depth anything V2. In: Advances in Neural Information Processing Systems (NeurIPS). Red Hook, NY, USA: Curran Associates, Inc.; 2024. p. 21875–911.
67. Li X, Kuang P. 3D-VRVT: 3D voxel reconstruction from a single image with vision transformer. In: 2021 International Conference on Culture-Oriented Science & Technology (ICCST). Piscataway, NJ, USA: IEEE; 2021. p. 343–8. doi:10.1109/iccst53801.2021.00078.
68. Yu J, Yin W, Hu Z, Liu Y. 3D reconstruction for multi-view objects. Comput Electr Eng. 2023;106:108567. doi:10.1016/j.compeleceng.2022.108567.
69. Wu YL, Shuai HH, Tam ZR, Chiu HY. Gradient normalization for generative adversarial networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE; 2021. p. 6353–62. doi:10.1109/iccv48922.2021.00631.
70. Serin Z, Yüzgeç U, Karakuzu C. Pre-trained variational autoencoder approaches for generating 3D objects from 2D images. In: 2nd International Congress of Electrical and Computer Engineering. Cham, Switzerland: Springer; 2023. p. 87–101. doi:10.1007/978-3-031-52760-9_7.
71. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of Wasserstein GANs. In: Advances in Neural Information Processing Systems (NeurIPS). Red Hook, NY, USA: Curran Associates, Inc.; 2017. p. 5767–77.
72. Xu M. Towards generalized implementation of Wasserstein distance in GANs. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Palo Alto, CA, USA: AAAI Press; 2021. p. 10514–22. doi:10.1609/aaai.v35i12.17258.
73. Kingma DP, Welling M. An introduction to variational autoencoders. Found Trends Mach Learn. 2019;12(4):307–92. doi:10.1561/2200000056.
74. Liu K, Qiu G. Lipschitz constrained GANs via boundedness and continuity. Neural Comput Appl. 2020;32(24):18271–83. doi:10.1007/s00521-020-04954-z.
75. Chang AX, Funkhouser T, Guibas L, Hanrahan P, Huang Q, Li Z, et al. ShapeNet: an information-rich 3D model repository. arXiv:1512.03012. 2015.
76. Wang N, Zhang Y, Li Z, Fu Y, Liu W, Jiang YG. Pixel2Mesh: generating 3D mesh models from single RGB images. In: The European Conference on Computer Vision (ECCV). Cham, Switzerland: Springer; 2018. p. 55–71. doi:10.1007/978-3-030-01252-6_4.