electronics

MDPI

Article

WGA-SWIN: Efficient Multi-View 3D Object Reconstruction Using Window Grouping Attention in Swin Transformer

Sheikh Sohan Mamun , Shengbing Ren *, MD Youshuf Khan Rakib and Galana Fekadu Asafa

School of Computer Science and Engineering, Central South University, Changsha 410083, China; sohanmamun.cst@csu.edu.cn (S.S.M.); khanushuf4619@csu.edu.cn (M.Y.K.R.); galana911@csu.edu.cn (G.F.A.)
* Correspondence: rsb@csu.edu.cn

Abstract: Multi-view 3D reconstruction aims to discover 3D characteristics based on visual information captured across multiple viewpoints. Transformer networks have shown remarkable success in various computer vision tasks, including multi-view 3D reconstruction. However, the reconstruction of accurate 3D shapes faces challenges when trying to efficiently extract and merge features across views. The existing frameworks struggled to capture the subtle relationships between the views, resulting in a poor reconstruction. To address this issue, we present a new framework, WGA-SWIN, for 3D reconstruction using multi-view objects. Our method introduces a Window Grouping Attention (WGA) mechanism that uses group tokens from different views for each window attention operation, enabling efficient inter-view and intra-view feature extraction. Diversity among various groups in a model contributes to the richness of feature learning, which results in advanced and dependable feature learning, resulting in more comprehensive and robust representations. Within the encoder swin transformer blocks, we integrated WGA to utilize both hierarchical design and shifted window attention mechanisms for efficient multi-view feature extraction. In addition, we developed a progressive hierarchical decoder that combines swin transformer blocks with 3D convolutions to utilize voxel representation, resulting in a high resolution for obtaining high-quality 3D reconstructions with fine structural details. The experimental results on the benchmark datasets ShapeNet and Pix3D demonstrate that our work achieves state-of-the-art (SOTA) performance, outperforming existing methods in both single-view and multi-view 3D reconstruction, beyond the capabilities of current technologies. We lead by 0.95% and 1.07% in both IoU and F-Scores respectively, which demonstrates the robustness of our method.

Keywords: swin transformer; window grouping attention; progressive hierarchical decoder; multi-view 3D reconstruction

check for updates

Academic Editor: Daniel Morris

Received: 10 February 2025
Revised: 14 April 2025
Accepted: 15 April 2025
Published: 17 April 2025

Citation: Mamun, S.S.; Ren, S.; Rakib, M.Y.K.; Asafa, G.F. WGA-SWIN: Efficient Multi-View 3D Object Reconstruction Using Window Grouping Attention in Swin Transformer. *Electronics* **2025**, *14*, 1619. https://doi.org/10.3390/electronics14081619

1. Introduction

Multi-view 3D object reconstruction is an application of modern computer vision that has wide applications in many fields such as augmented reality, cultural heritage protection, autonomous driving, robotics, and medical imaging. The most important challenge in 3D reconstruction is the extraction of accurate and efficient features of an object from multiple views in order to create realistic and detailed models of objects with 3D representation [1]. Traditional reconstruction methods, such as SfM (Structure from Motion) [2] and SLAM (Simultaneous Localization and Mapping) [3], have shown difficulty in balancing accurate reconstruction, especially in real-time and large-scale applications.

Nowadays, several deep learning methods have shifted towards recurrent neural networks (RNNs) [4,5] and convolutional neural networks (CNNs) [6–10], which have demonstrated significant success in feature extraction and spatial understanding of 3D reconstruction tasks. However, these methods experience difficulties in optimizing long-range dependencies while maintaining effective information exchange between multiple views during the encoding process. In contrast, the introduction of transformer-based [11–15] models has attracted attention for demonstrating the capability to capture long-range dependencies among relationships between input tokens, which makes these models particularly effective for visual understanding of spatial data across multiple views. Vision Transformers (ViTs) [16] are demonstrated as powerful tools for capturing global information through their attention mechanism, which splits images into fixed-size patches and analyzes them sequentially. When applied to multi-view 3D reconstruction tasks, ViTs face difficulties because of their fixed patch size along with performance inefficiencies as well as the heavy amount of view inputs [17].

Recently, swin transformers [18–20] have shown great promise in addressing many of these processing challenges by combining a hierarchical design with shifted window attention in the field of computer vision. Through their hierarchical design, swin transformers can efficiently process large-scale data while maintaining the ability to capture long-range dependencies. Unlike ViTs, swin transformers divide the image into local windows for attention, increasing the receptive field along the layer path, which makes them very well-suited for processing multi-view data in 3D reconstruction. Despite these advantages, it remains a challenge to effectively handle the intricacy of multi-view 3D data, especially when views increase and the model has to handle a large number of token interactions between views.

To address this issue, the proposed WGA-SWIN-based 3D reconstruction method introduces a novel encoder framework, which integrates the Window Grouping Attention (WGA) mechanism into the swin transformer architecture. The WGA mechanism segments tokens from each window into groups based on semantic importance, allowing separate attention operations within each group. This effectively captures intra- and inter-view relationships, increasing efficiency while preserving a global feature representation. This strategy allows the model to focus on the most relevant features while maintaining a large representation. By encoding the WGA mechanism into the swin transformer architecture in the encoder, the proposed grouping technique evaluates the importance of tokens from each different window based on their attention value and their hierarchical design and shifts window attention to extract features from multiple views efficiently, thus extracting local and global dependencies for high-quality 3D reconstruction.

For our task, we developed a progressive hierarchical decoder that integrates transposed convolution into the swin transformer blocks, which progressively up-sampled voxel representations to ensure high-resolution and detailed 3D reconstructions. This decoder maintains the fine-grained details required for high-quality reconstructions while balancing the reconstruction efficiency during the up-sampling process. The proposed method addresses a significant advancement to increase the multi-view 3D reconstruction capabilities while addressing the main issue related to feature richness and representation efficiency. The experimental results on the ShapeNet and Pix3D datasets demonstrated that our model established a SOTA performance in the area of 3D object reconstruction.

The contributions of our work are as follows:

(1) We present a Window Grouping Attention (WGA) mechanism, which can segment tokens into groups from individual window attentions and efficiently establish both inter-view and intra-view feature extraction. A novel encoder is introduced for operating multi-view inputs by integrating WGA into a swin transformer architecture to enhance 3D reconstruction accuracy.

(2) We introduce a progressive hierarchical decoder to integrate a 3D CNN and a swin transformer block to address the dependencies of working on the comparatively high-resolution voxel that can enhance reconstruction capability.

(3) The experimental results on the datasets ShapeNet and Pix3D represent that our work outperforms other SOTA methods in both multi-view and single-view 3D object reconstruction.

The main sections of this article are organized as follows. In Section 2, we present a review of related work, including multi-view 3D reconstruction, especially transformer-based 3D reconstruction methods. Section 3 introduces the proposed WGA-SWIN framework, including the Window Grouping Attention (WGA) mechanism, an encoder based on a swin transformer, and a progressive hierarchical decoder. To demonstrate the effectiveness of our proposed approach, extensive experiments, results, and ablation analysis are conducted within Section 4. Finally, the conclusion and suggestions for further research are presented in Section 5.

2. Related Work

2.1. Multi-View 3D Reconstruction

Recent years have seen significant advancements in multi-view 3D object reconstruction tasks according to deep learning techniques, of which transformer-based models are particularly effective [1]. By establishing features from multiple objects, traditional methods, like SfM [2] and SLAM [3], have been frequently employed to reconstruct 3D scenes. For multi-view tasks, deep learning algorithms, such as RNNs and CNNs, can improve feature extraction and fusion. While RNN-based methods 3D-R2N2 [5] and LSM [4] are inefficient and computationally expensive, CNN-based methods, like 3D-DensityNet [8], use maximum pooling but lack view connectivity. The CNN approaches such as Attsets [6], GARNet [10], and Pix2Vox [7] make use of attention mechanisms to preserve the relationship between different views, but these mechanisms are often simple; as a result, they cannot model complex multi-view relationships. Sequential transformer approaches, such as EVolT [12] and LegoFormer [15], utilize CNN backbone models to compress tokens across views, yet this reduction reduces the expressiveness of single-view representations. As a result, transformer approaches, such as EVolT [12] and LegoFormer [15], use the CNN backbone to compress the number of tokens per view, but this compression reduces the number of views represented by individual views. On the other hand, the cross-view processing of the 3D-RETR [14] model operates on isolated views with transformers until the feature proposition combines features but falls short in preserving essential inter-view dependencies essential to multi-view 3D reconstruction. UMIFormer [13] and LRGT [11] attempt to address these challenges while decoupling feature extraction within and between views; adding additional modules to infer relationships between views increases the computational complexity of the model. However, these methods struggle to handle complex and large-scale environments, especially when feature matching is ambiguous or computationally efficient, which is a limiting factor. Despite promising results, these methods fail to capture strong inter-view relationships because they handle all image tokens equally for each view.

Unlike multi-view fusion methods, 3D Gauss [21] and NeRF [22] focus on improving 3D reconstruction efficiency and detail. Three-dimensional Gauss uses Gaussian splatting for better scalability and uncertainty handling in large scenes. NeRF [22] generates detailed 3D models from sparse 2D images, capturing complex light interactions and volumetric scenes, and it has been extended to handle dynamic scenes to improve efficiency and accuracy in 3D reconstruction.

2.2. Group Window Attention

In this research, our main focus is on Window Grouping Attention-based 3D object reconstruction. Recently, some researchers have proposed the token grouping technique [23] to aggregate similar tokens [17] from different views and improve the accuracy of 3D object reconstruction. However, the introduction of ViT [16] represents a breakthrough in how computer vision handles image processing. ViT groups image information into tokens and apply self-attention to understand global dependencies. Demonstrating the application of transformers in multi-view analysis to capture spatial and temporal relationships has been explored by employing STTN [24] for video retrieval systems. STTN utilizes full-range attention in both spatial and temporal dimensions [24], and it is while effective for video, it faces challenges in efficiently handling large datasets. To solve this problem, DSTT [25] employs short-range grouping attention and increases local connectivity within each view, enhancing feature extraction for temporal coherence tasks to increase computing efficiency. However, it struggles to build long-range dependencies between views and cannot effectively handle interactions across multiple views, which is a significant problem in multi-view 3D reconstruction tasks. LRGT [11] uses long-range attention, which merges tokens from fixed positions in the images, attempting to capture long-range dependencies between views. It improves inter-view correlations but faces difficulties in grouping semantically similar tokens between multiple views and lacks a mechanism for handling intra-view relationships, which are essential for detailed 3D reconstruction, so attention-guided fusion [26] is introduced to enhance multi-view feature integration. Recently, the swin transformer [20,27] has introduced window-based attention, such as R3D-Swin [19], where an image is divided into non-overlapping windows, and attention is computed within individual windows. To capture long-range dependencies between windows, the swin transformer employs a window-shifting technique [28]. Compared to traditional transformers, which compute attention for the entire image, this approach reduces computational complexity. In addition, the fixed window structure still limits the model's ability to effectively capture complex relationships between views in multi-view 3D reconstruction tasks.

As a result, we propose the Window Grouping Attention (WGA) mechanism, which segments tokens within each window based on semantic importance, allowing separate attention operations within each group. This approach enables efficient attention processing while preserving critical intra- and inter-view relationships. By focusing on the most relevant features and flexible attention manipulation, WGA addresses the difficulties of working with large and complex datasets and enhances the performance without compromising the global feature representation necessary for high-quality 3D reconstruction.

3. Methods

As represented in Figure 1, our proposed WGA-SWIN is based on a transformer-based architecture that has been modified with a novel module that enhanced the swin transformer and window token grouping. The importance behind WGA-SWIN lies in extracting information-rich inter-view features with high-quality 3D object reconstruction. The proposed method is capable of processing a variety number of view images with dimensions of 224 × 224 × 3, denoted as the image set $I = \{I_1, I_2, \ldots, I_N\}$, and generating the corresponding voxel representation V with dimensions of 32^3. To enhance feature integration across views, we introduce a Window Grouping Attention (WGA) mechanism, which segments tokens into groups for independent window attention operations.

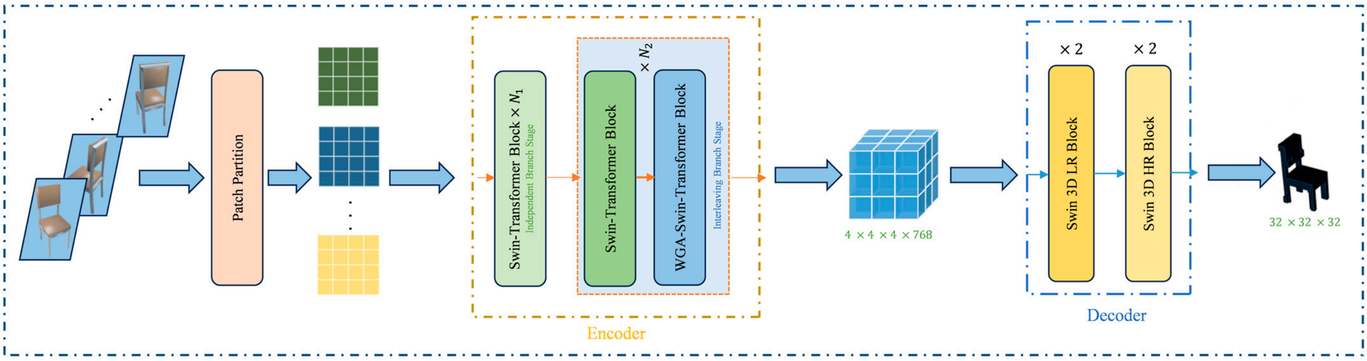


Figure 1. The overall architecture of our proposed WGA-SWIN.

The detailed process for the encoders is described in Section 3.1. After feature extraction, the proposed method integrates a progressive hierarchical decoder based on swin transformer blocks with a 3D CNN. The decoder progressively upscales the feature map into a voxel grid, preserving fine-grained details and ensuring visualization efficiency, as described in Section 3.2. The voxel representation V is constructed through successive decoding operations. The overall transformation process is summarized as follows:

$$V = \text{WGA-SWIN}(I) = De(En(I)) \tag{1}$$

where En and De represent the processes of feature extraction by the encoder, multi-view feature processing through the WGA-SWIN mechanism, and the generation of the 3D output V through the decoder, respectively.

3.1. Encoder

The encoder architecture is designed with the architecture swin transformer [18] and integrates the proposed Window Grouping Attention (WGA) mechanism. It has been split into two parts: an independent and an interleaving branch stage. The initial stage consists of an N_1 standard swin transformer unit shown in Figure 2a. The second stage alternates between intra-view and inter-view feature processing, utilizing N_2 blocks, including WGA, to establish correlations between views, as shown in Figure 2b. After that, the latter utilizes swin transformer blocks with the WGA mechanism and ensures efficient feature aggregation by grouping tokens into windows for localized attention operations, while the Inter-View Feature Signatures (IFSs) [11] enhance the encoder's ability to distinguish features across views. The overall architecture is shown in Figure 2. In the end, we use the similar-token merger [29] to decrease the tokens of all branches of the specified size.

3.1.1. Independent Branch Stage

At this stage, each input view is partitioned into independent windows, and the tokens from those windows are embedded linearly trainable projections into a feature space. The system processes these tokens through N_1 swin transformer blocks shown in Figure 2a that establish self-attention between tokens within local windows. The attention mechanism enhances computation efficiency by preventing attention to tokens between each window while capturing fine-grained spatial dependencies. The self-attention operation across a single window is defined as follows:

$$\text{Attention}_{(\text{window})}(Q,\, K,\, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

where $Q, K, V \in \mathbb{R}^{M_w \times D}$ represent queries, keys, and value matrices obtained from M_w the local window of token embeddings and $d_k = D$ is the number of feature dimensions used for scaling. The output of this phase is a spatial feature (window by window) that contains the k-th view. The output of this stage is a spatial feature map for each view, representing the internal structure of the object within that specific view. This stage ensures that the encoder captures detailed representations of the individual views before combining them in the subsequent stage.

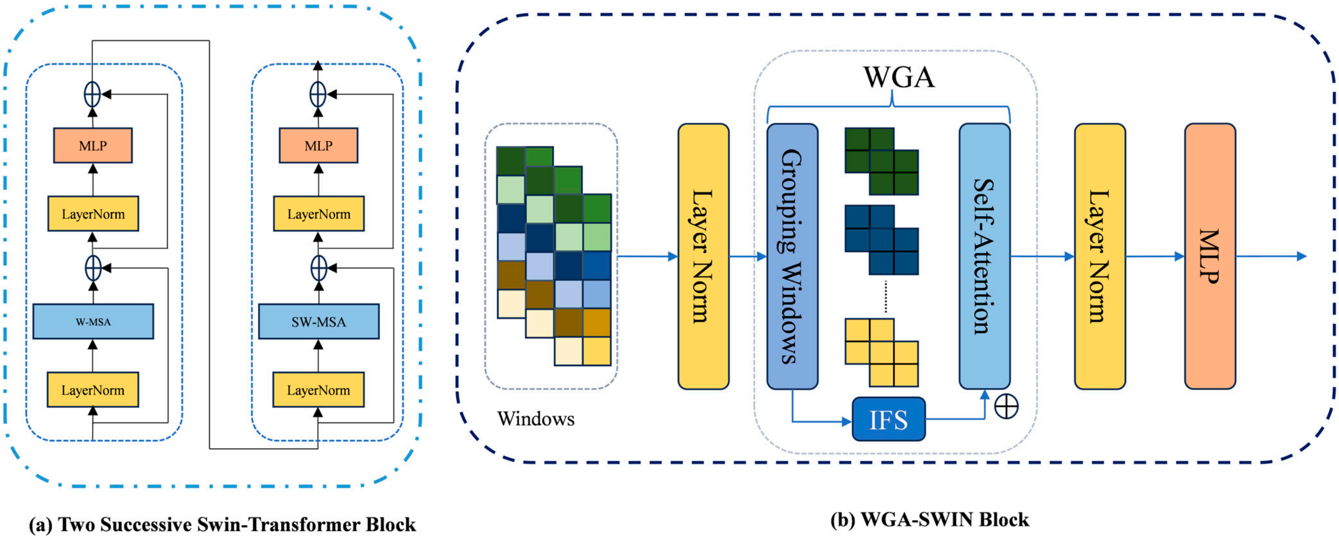


Figure 2. Illustration architecture of the encoder, (**a**) swin transformer block, and (**b**) Window Grouping Attention (WGA) block.

3.1.2. Interleaving Branch Stage

This interleaving branch stage alternates between using swin transformer blocks and Window Grouping Attention (WGA) for inter-view communication, as shown in Figure 2b. This stage ensures the efficient integration of spatial and relational information across multiple views.

This section introduces the Window Grouping Attention (WGA) mechanism that effectively captures the relationships between views. For a given input feature view $X_i \in \mathbb{R}^{P \times D}$, where the number of tokens is P, the feature dimension is D, and tokens are grouped into G groups, each group G_j is defined as follows:

$$G_j = \{x_{i,k} \mid i = 1, 2, \ldots, N;\ k \in W_j\} \tag{3}$$

where $x_{i,k}$ denotes the k-th token from the i-th view and W_j specifies the index of tokens in the j-th spatial window. This ensures that grouping tokens in G_j comes from the same spatial window in all views, promoting local attention by maintaining inter-view consistency, which is key for capturing significant spatial relationships and ensuring that features from each view contribute to the final aggregated representation.

Within each group G_j, tokens engaged independent self-attention, and they are computed as follows:

$$\text{Attention}_{\text{j}_{(\text{window})}}(Q_j, K_j, V_j) = \text{Softmax}\left(\frac{Q_j {K_j}^T}{\sqrt{d_k}}\right) V_j \tag{4}$$

where $Q_j, K_j, V_j \in \mathbb{R}^{M_g \times D}$ represent the queries, keys, and value matrices obtained from the grouped tokens and the number of tokens from each group M_g. By reducing self-attention between grouped tokens, WGA establishes efficient computation while preserving long-range relationships between different views.

After applying the self-attention operation, all groups are sent for separate processing before their outputs are combined into an inter-view feature representation. This divide-and-conquer approach allows WGA to focus on token correlations between spatial windows across views, effectively understanding intra-view details alongside inter-view dependencies.

Furthermore, to improve inter-view associations, Inter-View Feature Signatures (IFSs) [11] have been implemented. The IFS assigns view-specific encodings to the tokens, which ensures that the model can distinguish tokens originating from different views. Specifically, for each token $x_{i,j}$ in the group G_j, the feature signature is defined as $f_{i,j} = \text{IFS}\left(\phi\left(x_{i,j}\right)\right)$, where $\phi\left(x_{i,j}\right)$ denotes the trainable projection. The feature signatures for all views are expressed as $f_j = \left\{f_{1,j},\ f_{2,j},\ \ldots,\ f_{N,j}\right\}$, where N denotes the number of views. When the IFS is added to the output of WGA, the following occurs:

$$\text{Attention}_{\text{j}_{(\text{window})}}\left(Q_j,\ K_j,\ V_j\right) = \text{Softmax}\left(\frac{Q_j {K_j}^T}{\sqrt{d_k}}\right)V_j + f_j \tag{5}$$

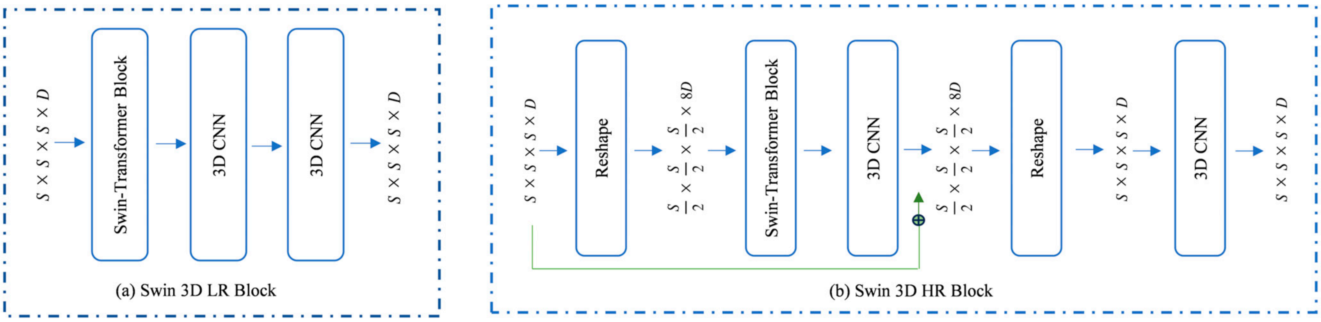
where the addition of the IFS in WGA allows the model to distinguish between tokens from different views, enriching feature representation with view-specific information. This enhances the model's ability to maintain consistency across multiple views and improves the accuracy of the feature aggregation process.

The interleaving branch stage combines WGA with the view-specific differentiation of the IFS to produce a characterized, diverse, and consistent representation of features across multiple views. This allows the encoder to efficiently handle multi-view inputs and provide rich feature maps for critical tasks, such as 3D reconstruction.

3.2. Decoder

In our study, the previous transformer-based decoders for 3D voxel reconstruction, attention mechanisms, have been employed to reconstruct voxel grids, and these approaches often generate low-resolution (LR) features and up-sample them rapidly to the final target resolution in the later stages of the decoder. Although this results in an incomplete reconstruction of delicate details, typical global attention [29] has difficulty handling the enormous number of tokens associated with high-resolution (HR) features. To solve this restriction, we designed a progressive hierarchical decoder that combines swin transformer blocks with 3D CNN [14] layers. This hybrid architecture progressively up-samples voxel representations, allowing the model to optimize both global and local features while preserving structural details.

Our proposed decoder progressively reconstructs the voxel grid from 4^3 to 32^3 resolution using Swin 3D LR Blocks and Swin 3D HR Blocks, as illustrated in Figure 3. Each block utilized the benefits of 3D convolution for local feature extraction and swin transformer blocks for long-range dependency modeling. Together, these components work to improve voxel representations across multiple resolutions, assuring efficiency and high-fidelity reconstructions.


$S \times S \times S \times D$
Swin-Transformer Block
3D CNN
3D CNN
$S \times S \times S \times D$
$S \times S \times S \times D$
Reshape
$\frac{S}{2} \times \frac{S}{2} \times \frac{S}{2} \times 8D$
Swin-Transformer Block
3D CNN
$\frac{S}{2} \times \frac{S}{2} \times \frac{S}{2} \times 8D$
Reshape
$S \times S \times S \times D$
3D CNN
$S \times S \times S \times D$

(a) Swin 3D LR Block

(b) Swin 3D HR Block

Figure 3. An illustration of the decoder, (**a**) Swin 3D LR Block, and (**b**) Swin 3D HR Block.

The structure of the Swin 3D LR Block easily processes the voxel features at a low resolution to extract coarse-level information and up-samples them for subsequent refinement. As shown in Figure 3a, given an input feature map $X \in \mathbb{R}^{s \times s \times s \times D}$, the block begins by applying a swin transformer block, which operates on non-overlapping spatial windows. Within each window, window-based attention (W-MSA) is computed independently, enabling the model to capture both local and global dependencies effectively. Next, the output of the swin transformer block then uses a 3D convolutional layer to refine local features and up-sample the feature map using a 3D transposed convolution (3D CNN). While maintaining coarse-level details, this process recovers the spatial resolution. The Swin 3D LR Block maintains an output feature map with a similar resolution $S \times S \times S \times D$, showing that each global context and local feature is captured for future enhancement.

The Swin 3D HR Block further refines voxel features at a higher resolution to preserve fine-grained spatial details, which are essential for accurate 3D reconstruction. As shown in Figure 3b, the input feature map $X \in \mathbb{R}^{s \times s \times s \times D}$ is first reshaped and downsampled to $\frac{s}{2} \times \frac{s}{2} \times \frac{s}{2} \times 8D$, reducing the spatial resolution while increasing the feature dimension. As a result, this increased the reconstruction efficiency and enabled the swin transformer block to process fewer tokens while capturing detailed spatial information. After that, the down-sampled features are passed through a swin transformer block, where window-based attention (W-MSA) is again applied to simulate both local and global dependencies within each window. To follow this, a 3D convolutional layer is applied to refine the voxel features further. After that, the output is then reshaped to the original resolution $S \times S \times S \times D$. In the end, we keep the skip connections between the input and the output of the block optimized in order to preserve the fine-grained spatial details. Later, the skip features are combined with the up-sampled features utilizing element-wise augmentation to ensure that both low-level details and high-level contextual information are preserved in the voxel map reconstruction. Finally, the modified voxel representation is then generated by processing the combined information through a 3D convolutional layer. Later, the swin transformer-based decoder's hierarchical structure efficiently incorporates W-MSA and 3D convolution operations. We assume that the Swin 3D LR Blocks will extract coarse features, while the Swin 3D HR Blocks will refine these features with skip connections to allow for an accurate and efficient reconstruction of the 3D voxel shape.

3.3. Loss Function

To optimize [14] the reconstruction accuracy of our 3D voxel shape, we follow Dice Loss [30] as a loss function to train the model, which has proven the ability to handle highly unbalanced voxel occupancy. Dice Loss is well suited for voxel-based reconstruction tasks where empty spaces dominate to calculate between the ground truth reconstruct voxel grids overlap. Dice Loss is formulated as follows:

$$L = 1 - \frac{\sum_{i=1}^{32^3} w_i s_i}{\sum_{i=1}^{32^3} w_i + s_i} - \frac{\sum_{i=1}^{32^3} (1-w_i)(1-s_i)}{\sum_{i=1}^{32^3} 2 - w_i - s_i} \tag{6}$$

where w_i and s_i reflect the confidence values of the i-th voxel grid in the ground truth and reconstructed volume, respectively.

4. Experimental Analysis

4.1. Dataset

In order to evaluate our proposed approach in this paper, the experiments are implemented on two datasets, ShapeNet [31] and Pix3D [32]. Following previous works [5,11], we utilize a subset of ShapeNet to train our model, which consists of 43,783 objects from 13 categories. Where providing an extensive dataset for voxel-based reconstruction, each object is rendered from 24 different viewpoints. For our work implementation, we use 30,642 models for training, 4371 for validation, and 8770 for testing. Each 3D model is rendered into 24 different viewpoint images of 137 × 137 with ground truth voxel grids provided at resolutions up to 32^3. On the other hand, to evaluate the real-world data of our model, we use the dataset Pix3D [32], which includes 2894 untouched and unconcluded chair categories of images as follows [14,19].

4.2. Evaluation Metrics

To evaluate our proposed method, we adopt the Intersection over Union (IoU) and F-Score@1% to follow similar evaluation methods of previous works [14]. They measure the accuracy of predicted voxel grids and reconstructed 3D surfaces, where higher values of the IoU and F-Score@1% indicate good reconstruction quality.

Intersection over Union (IoU) is used to evaluate the ground truth and reconstruct voxel grid overlap. It is formulated as follows:

$$IoU = \frac{\sum_{(i,\, j,\, k)} I\left(\hat{p}(i,\, j,\, k) > t\right) I(p(i,\, j,\, k))}{\sum_{(i,\, j,\, k)} I\left[I\left(\hat{p}(i,\, j,\, k) > t\right) + I(p(i,\, j,\, k))\right]} \tag{7}$$

where $\hat{p}(i,\, j,\, k)$ and $(p(i,\, j,\, k))$ represent the ground truth and predicted utilization probability at the voxel $(i,\, j,\, k)$ position, respectively. An indicator function $I(.)$, and the voxelization threshold denotes t. In the end, IoU quantifies how precisely the ground truth aligns with the reconstructed voxel grids.

F-Score@1% is used to measure the precision and recall of ground truth and predicted point clouds to evaluate the reconstruction quality of 3D surfaces [33]. It is formulated as follows:

$$F - Score(d) = \frac{2P(d)R(d)}{P(d) + R(d)} \tag{8}$$

where $P(d)$ and R(d) represent the precision and recall for a distance threshold d within the ground truth and the predicted points. These are calculated as follows:

$$P(d) = \frac{1}{n_{\mathcal{R}}} \sum_{r \in \mathcal{R}} [min_{g \in \mathcal{G}} \parallel r - g \parallel < d] \tag{9}$$

$$R(d) = \frac{1}{n_{\mathcal{G}}} \sum_{r \in \mathcal{G}} [min_{r \in \mathcal{R}} \parallel g - r \parallel < d] \tag{10}$$

where $\mathcal{G}$ and $\mathcal{R}$ are the ground truth and reconstructed point clouds, $n_{\mathcal{G}}$ and $n_{\mathcal{R}}$ are the number of points in G and R, respectively, and $[.]$ is the Iverson bracket. For the voxel representations, the object surface is generated using the Marching Cubes algorithm, and 8192 points are taken from the surface in order to measure the F-Score. The F-Score@1% refers value of the F-Score when the distance threshold d is set to 1%.

4.3. Experimental Setup

For a fair comparison with previous works [11,13], our model is evaluated to initialize the encoder's network parameters using the pre-training model Swin-B [27] and allow the model parameters to be updated and optimized during the training process. In the WGA swin transformer, block G is defined as 49 in WGA. The encoder contains 12 swin transformer blocks, where N_1 = 6 and N_2 = 3. Additionally, in the decoder, the number of window tokens n is set to 64, and the decoder is built with Swin 3D LR Blocks and Swin 3D HR Blocks, which progressively up-samples voxel features from a low-resolution grid 4^3 to a high-resolution grid 32^3 with skip connections used to maintain fine-grained spatial details during reconstruction. To validate the reconstruction performance, we provide two models, namely, WGA-SWIN and WGA-SWIN+, which share the same network architecture but are different in the number of views [11] used for training (three views for WGA-SWIN and eight views for WGA-SWIN+). The experimental code is implemented in PyTorch 2.4.1 and trained on two NVIDIA RTX 2080 Ti GPUs with 24 GB memory. All input images are resized to 224 × 224, and 3D voxel grids are normalized before processing. Following the previous works [13], for our WGA-SWIN and WGA-SWIN+, the training process uses the AdamW [34] optimizer, with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The threshold t for voxel binarization is set to 0.5 for WGA-SWIN and 0.4 for WGA-SWIN+. The training batch size is 16, and the model is trained for a total of 150 epochs. The primary learning rate is set to 0.0001 and decreased by a ratio of 0.1 after the 70th and 100th epochs. The training batch size is 16, and the model is trained for a total of 150 epochs. To evaluate reconstruction performance, we use Dice Loss as the loss function to handle imbalanced voxel occupancy. During inference, both WGA-SWIN and WGA-SWIN+ can adapt to an arbitrary number of input views, with performance evaluated using IoU and F-Score@1% metrics.

4.4. Results and Discussion

In this section, we present a series of tests conducted to evaluate the effectiveness of our proposed framework for multi-view 3D reconstruction. The results aim to answer the following research questions (RQs).

4.4.1. Quantitative Results

RQ1: *How well does the proposed framework, WGA-SWIN, outperform the SOTA approach on ShapeNet in the task of multi-view 3D reconstruction, and how does its performance rely on different input views?*

Compared to the state-of-the-art (SOTA) methods [6,10,11,13,14] on the ShapeNet test dataset, our proposed WGA-SWIN framework exhibits superior performance and presents the results in the form of the IoU and F-Score@1%. The test results, shown in Tables 1 and 2, show that our WGA-SWIN and WGA-SWIN+ significantly outperform the existing SOTA methods on different numbers of input views, which range from 1 to 20 views and are consistent with standard benchmarks in multi-view reconstruction tasks, where they represent the best compromise for reconstruction accuracy. Both WGA-SWIN and WGA-SWIN+ consistently outperform existing methods including LRGT [11], UMIFormer [13], and GARNet [10] for high-quality 3D voxel grid reconstruction. On the other hand, for single-view reconstruction, our WGA-SWIN achieves 0.7032 IoU, which is better than the second-best LRGT's 0.6962 IoU. For multi-view reconstruction, when trained with 20 input views, our WGA-SWIN+ achieved a SOTA performance of 0.8017 IoU, which is 0.95% better than LRGT+. Furthermore, the F-Score@1% performance improvement is even more noticeable, further highlighting the ability of WGA-SWIN and WGA-SWIN+ to reconstruct objects with superior accuracy and detail.

Table 1. Evaluations of performance comparison on ShapeNet using IoU for the multi-view 3D reconstruction approach. The best results are highlighted in bold.

Methods	**1 View**	**2 Views**	**3 Views**	**4 Views**	**5 Views**	**8 Views**	**12 Views**	**16 Views**	**20 Views**
3D-R2N2 [5]	0.560	0.603	0.617	0.625	0.634	0.635	0.636	0.636	0.636
AttSets [6]	0.642	0.662	0.670	0.675	0.677	0.685	0.688	0.692	0.693
Pix2Vox++ [9]	0.670	0.695	0.704	0.708	0.711	0.715	0.717	0.718	0.719
GARNet [10]	0.673	0.705	0.716	0.722	0.726	0.731	0.734	0.736	0.737
GARNet++	0.655	0.696	0.712	0.719	0.725	0.733	0.737	0.740	0.742
EVolT [12]	-	-	-	0.609	-	0.698	0.720	0.729	0.735
LegoFormer [15]	0.519	0.644	0.679	0.694	0.703	0.713	0.717	0.719	0.721
3D-RETR [14]	0.674	0.707	0.716	0.720	0.723	0.727	0.729	0.730	0.731
UMIFormer [13]	0.6802	0.7384	0.7518	0.7573	0.7612	0.7661	0.7682	0.7696	0.7702
UMIFormer+	0.5672	0.7115	0.7447	0.7588	0.7681	0.7790	0.7843	0.7873	0.7886
LRGT [11]	0.6962	0.7462	0.7590	0.7653	0.7692	0.7744	0.7766	0.7781	0.7786
LRGT+	0.5847	0.7145	0.7476	0.7625	0.7719	0.7833	0.7888	0.7912	0.7922
WGA-SWIN	**0.7032**	**0.7487**	**0.7636**	**0.7732**	0.7748	0.7786	0.7808	0.7819	0.7827
WGA-SWIN+	0.5976	0.7217	0.7563	0.7712	**0.7814**	**0.7882**	**0.7955**	**0.7996**	**0.8017**

Table 2. Evaluations of performance comparison on ShapeNet using F-Score@1% for the multi-view 3D reconstruction approach. The best results are highlighted in bold.

Methods	**1 View**	**2 Views**	**3 Views**	**4 Views**	**5 Views**	**8 Views**	**12 Views**	**16 Views**	**20 Views**
3D-R2N2 [5]	0.351	0.372	0.372	0.378	0.382	0.383	0.382	0.382	0.383
AttSets [6]	0.395	0.418	0.426	0.430	0.432	0.444	0.445	0.447	0.448
Pix2Vox++ [9]	0.436	0.452	0.455	0.457	0.458	0.459	0.460	0.461	0.462
GARNet [10]	0.418	0.455	0.468	0.475	0.479	0.486	0.489	0.491	0.492
GARNet++	0.399	0.446	0.465	0.475	0.481	0.491	0.498	0.501	0.504
EVolT [12]	-	-	-	0.358	-	0.448	0.475	0.486	0.492
LegoFormer [15]	0.282	0.392	0.428	0.444	0.453	0.464	0.470	0.472	0.472
3D-RETR [14]	-	-	-	-	-	-	-	-	-
UMIFormer [13]	0.4281	0.4919	0.5067	0.5127	0.5168	0.5213	0.5232	0.5245	0.5251
UMIFormer+	0.3177	0.4568	0.4947	0.5104	0.5216	0.5348	0.5415	0.5451	0.5466
LRGT [11]	0.4461	0.5005	0.5148	0.5214	0.5257	0.5311	0.5337	0.5347	0.5353
LRGT+	0.3378	0.4618	0.4989	0.5161	0.5271	0.5403	0.5467	0.5497	0.5510
WGA-SWIN	**0.4596**	**0.5095**	**0.5261**	**0.5312**	0.5354	0.5408	0.5427	0.5442	0.5457
WGA-SWIN+	0.3492	0.4723	0.5109	0.5297	**0.5379**	**0.5492**	**0.5565**	**0.5594**	**0.5617**

Additionally, for the single-view reconstruction, our WGA-SWIN achieves 0.4536 F-Score@1%, which outperforms the second-best LRGT, with 0.4461 F-Score@1%. However, when trained with 20 input views, WGA-SWIN+ achieves a SOTA performance of 0.5608 F-Score@1%, which outperforms LRGT+ by 1.07%. These results demonstrate how the proposed methods enhance the model's ability to capture the multi-view dependencies with higher accuracy to reconstruct 3D voxel grids.

In addition, the performance improvement after four views is due to the framework design. The WGA process groups tokens from each window to ensure efficient inter-view relationships between views. This architectural adjustment allows the proposed framework to effectively integrate information from other views, resulting in continuous performance improvements. For example, both WGA-SWIN and WGA-SWIN+ show significant improvements over LRGT and other SOTA methods as the number of views rises, with the difference expanding as the view count increases. These results demonstrate how well our WGA-SWIN captures multi-view dependencies and provides accurate 3D reconstructions. For multi-view reconstruction, we emphasize how well WGA-SWIN and WGA-SWIN+ handle different numbers of input views.

4.4.2. Qualitative Results

RQ2: *How does the proposed framework, WGA-SWIN, achieve accurate and visually detailed 3D reconstructions compared to state-of-the-art methods on ShapeNet with different input view counts?*

In this evaluation, we compare our method, both WGA-SWIN and WGA-SWIN+, following SoTA works from [10,11,13] using different numbers of view inputs from the ShapeNet dataset. The reconstruction results visualized in Figure 4 highlight the effectiveness of the proposed WGA-SWIN framework in handling the 3D reconstruction of multi-view tasks with various quantities of input (5, 10, 15, and 20) views on various ShapeNet objects. However, WGA-SWIN and WGA-SWIN+ frequently outperform state-of-the-art methods such as LRGT, UMIFormer, and GARNet by capturing both global and local fine-grained object details. Even with only five input views, WGA-SWIN demonstrates the ability to reconstruct the overall shape and basic features of an object more efficiently than competing methods. In addition, by increasing the number of views to 10, 15, and 20, the framework shows significant improvements in retaining complex details that are often overlooked by other methods such as sharp edges, smooth surfaces, and fine structures. This advantage is due to the Window Grouping Attention mechanism, which efficiently captures inter-view relationships, and the progressive hierarchical decoder, which refines voxel grids while preserving structural integrity.

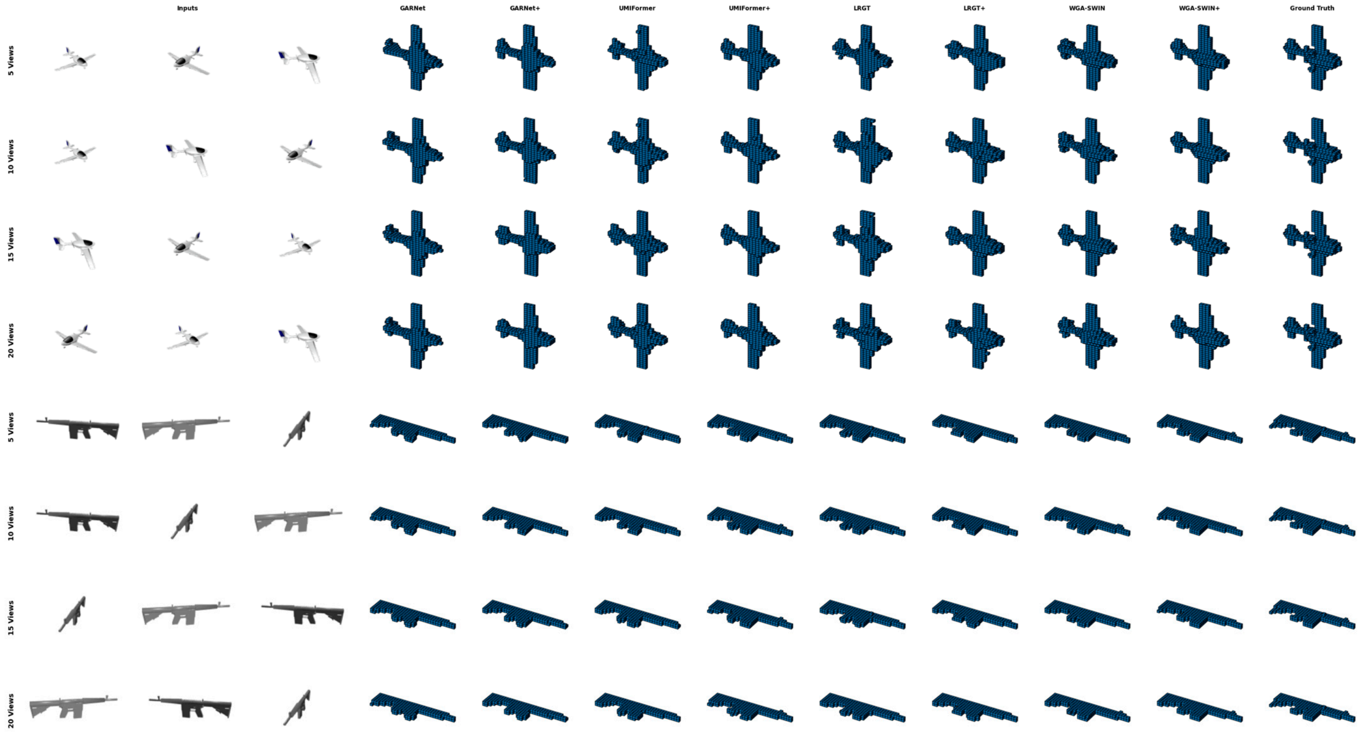


Figure 4. The results of multi-view 3D reconstruction on the test set of ShapeNet when facing 5, 10, 15, and 20 views as inputs.

To deal with objects that have complex geometries, such as airplanes and rifles, our proposed method, WGA-SWIN, excels in reconstructing key features such as heads, backrests, and tail fins with surprising accuracy. As the number of views increases, its ability to effectively integrate information from other views of the structure becomes more apparent, providing consistent structural accuracy and visual fidelity. However, when evaluated with 20 views, WGA-SWIN+ achieves superior reconstructions, which maintain fine-grained details and structural coherence better than any other approach to ensure its robustness and scalability. These results validate the adaptability and effectiveness of the proposed framework, demonstrate its ability to provide high-quality 3D reconstructions over a wide range of input views, and further establish it as a leading method for multi-view reconstruction tasks.

4.4.3. Evaluation of Real-World Objects

RQ3: *How well does the proposed method operate real-world objects with the Pix3D dataset?*

To evaluate the proposed WGA-SWIN method's performance in real-world image processing, we conduct experiments on the Pix3D dataset, which is a difficult benchmark for single-view 3D reconstruction. The dataset contains real-world images with complicated backgrounds, different lighting conditions, and occlusions. Following the setting in [7,9], the chair category from ShapeNetChairRFC is used to generate the training set, which uses 60 synthesized images generated for each object with randomly selected backgrounds from the Sun database [35,36].

For this evaluation, we compared our WGA-SWIN framework with available state-of-the-art methods, including UMIFormer [13] and LRGT [11], as shown in Table 3. Although Pix3D provides only a single-view input for each object, which slightly limits the performance of methods, like WGA-SWIN, designed for multi-view inputs, our framework still achieves the best performance in both the IoU and F- Score@1% metrics. However, WGA-SWIN outperforms all other approaches, with an IoU of 0.321 and an F-Score@1% of 0.147, compared to the closest competitor, UMIFormer, which scores 0.317 IoU and 0.142 F-Score@1%. This demonstrates how reliable our framework is in reconstructing real-world objects.

Table 3. Evaluation on Pix3D using IoU/F-Score@1%. The best result is highlighted in bold.

UMIFormer [13]	**LRGT [11]**	**WGA-SWIN**
0.317/0.142	0.302/0.129	**0.321/0.147**

Furthermore, the excellent performance can be attributed to the architectural innovations of WGA-SWIN, especially the progressive hierarchical decoder, which combines attention processes with convolutional layers. This design enables the model to capture complex structural details, preserve fine-grained features, and produce accurate reconstructions, even in single-view settings. As shown in Figure 5, WGA-SWIN excels in reconstructing fine details, such as the legs and backrest of a chair, which are often overlooked by competing methods. These results demonstrate that the proposed WGA-SWIN framework is not only effective in multi-view situations but also highly adaptable and competitive in real-world single-view reconstruction tasks. This adaptability and generalizability make it a strong candidate for 3D reconstruction in various application scenarios.

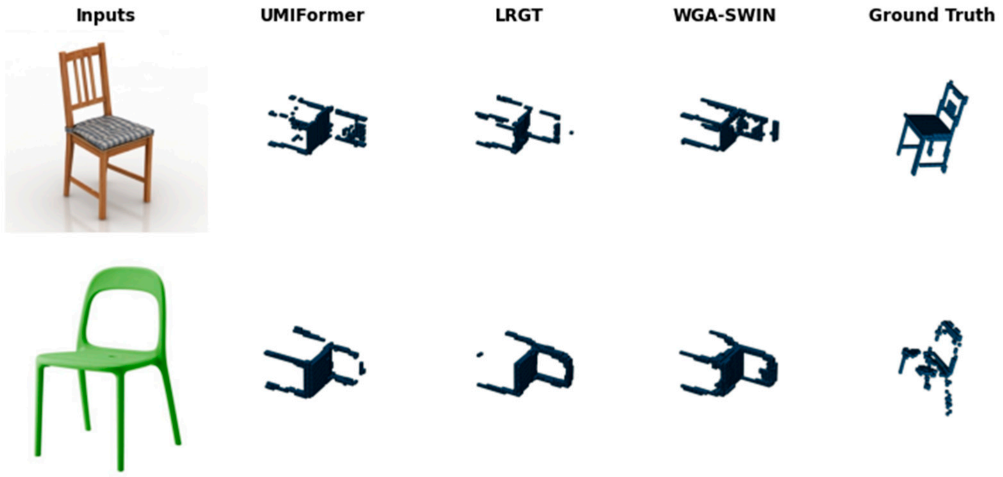


Figure 5. The results of single-view 3D reconstruction on the test set of Pix3D.

4.5. Ablation Study

In this section, we analyzed our encoder and decoder's ablation experiment results.

4.5.1. Effectiveness of Encoder

In order to compare our proposed Window Grouping Attention (WGA) mechanism in the swin transformer encoder by comparing it with the standard window-based attention (W-MSA), we conducted an ablation study to evaluate the impact of our model. The results, which are shown in Table 4, demonstrate how well WGA works to enhance reconstruction performance for different input view counts. This efficiency suffers extensively when WGA is replaced with W-MSA, which handles each view separately without inter-view communication, especially when the number of views increases. For example, for three input views, W-MSA has an IoU of 0.7562, which is significantly lower than the WGA-enabled configuration, which reaches 0.7636. However, when increasing the performance difference, the number of input views increases. Finally, WGA's IoU was 0.7819 at 16 views, while W-MSA's IoU fell to 0.7783. At 20 views, WGA's IoU was 0.7827, 0.33% higher than W-MSA's.

Table 4. An ablation experiment examining the impact of Window Grouping Attention (WGA) and window multi-head self-attention (W-MSA). Based on ShapeNet, IoU metrics are used to evaluate the experiments.

WGA	W-MSA	3 Views	5 Views	8 Views	12 Views	16 Views	20 Views
×	✓	0.7562	0.7679	0.7713	0.7758	0.7783	0.7790
✓	×	**0.7636**	**0.7748**	**0.7786**	**0.7808**	**0.7819**	**0.7827**

The table is marked with ✓ indicates presence of the feature and × indicates absence of the feature. The best results are highlighted in bold.

In contrast, these results show that WGA is essential for establishing inter-view communication, which improves the model's ability to represent inter-view dependencies. On the other hand, as the number of input views increases, a W-MSA that handles views independently will reach performance saturation. As an example of WGA's scalability, it significantly increases IoU performance across all view configurations while preserving model efficiency by focusing on grouped tokens. As a result of these findings, it can be determined that when dealing with a large number of input views, WGA plays an essential role in robust feature extraction in the encoder, thus enabling superior multi-view 3D reconstruction.

Additionally, Table 5 presents IoU scores for token grouping parameters G of 36, 49, and 64 across multiple input views. A grouping parameter of 36 consistently produces lower IoU scores, indicating that it fails to capture essential feature interactions, especially at lower view counts. On the other hand, 64 shows stable performance, but the gains over 49 are marginal, suggesting diminishing returns and unnecessary computational complexity. In contrast, 49 provides the best balance, achieving the highest reconstruction accuracy without excessive computational cost. Therefore, 49 was chosen for its optimal performance and efficiency across varying numbers of views.

Table 5. An ablation experiment examining the different grouping parameters G on ShapeNet using IoU score across views.

Grouping Parameters G	3 Views	5 Views	8 Views	12 Views	16 Views	20 Views
36	0.7592	0.7702	0.7744	0.7787	0.7798	0.7809
49	**0.7636**	**0.7748**	**0.7786**	**0.7808**	**0.7819**	**0.7827**
64	0.7612	0.7734	0.7772	0.7795	0.7808	0.7816

The best results are highlighted in bold.

4.5.2. Effectiveness of Decoder

As shown in Table 6, we evaluated the performance of our decoder with previous advanced works, including LegoFormer [15], 3D-RETR [14], and LRGT [11]. All experiments were performed using the same encoder as WGA-SWIN, and IoU was used for comparing the results on the ShapeNet dataset. The results demonstrate that our decoder performs better than others across every input view configuration. For instance, our decoder obtains an IoU of 0.7636 at three input views, which is higher than compared to 3D-RETR (0.7592) and LRGT (0.7581). However, our decoder achieves the highest IoU of 0.7827 at 20 input views, which is 0.68% higher than LegoFormer and 0.36% higher than LRGT.

Table 6. The performance comparison between our decoder and other previous works such as LegoFormer [15], 3D-RETR [14], and LRGT [11]. To handle the variables, the same encoder, WGA-SWIN, is applied in all experiments. Based on ShapeNet, IoU metrics are used to evaluate the experiments.

Decoder	3 Views	5 Views	8 Views	12 Views	16 Views	20 Views
LegoFormer [15]	0.7536	0.7643	0.7696	0.7729	0.7737	0.7751
3D-RETR [14]	0.7592	0.7685	0.7723	0.7775	0.7787	0.7796
LRGT [11]	0.7581	0.7687	0.7721	0.7766	0.7773	0.7785
WGA-SWIN (Ours)	**0.7636**	**0.7748**	**0.7786**	**0.7808**	**0.7819**	**0.7827**

The best results are highlighted in bold.

Finally, the superior performance of our decoder stems from its progressive design with SWIN 3D LR and SWIN 3D HR, which gradually increases voxel resolution through distributed swin transformer blocks at every stage. On the other hand, previous methods demonstrate different steps in applying simple structures, such as many fully connected layers. In addition, our approach allows better feature representation by combining window attention mechanisms for global refinement and convolution layers for local detail preservation. The results validate the effectiveness of our WGA-SWIN decoder, which achieves consistent improvements over prior methods, particularly with a large number of input views.

4.5.3. Inference Efficiency

As shown in Table 7, we compare the inference time of different methods, where the average inference time is measured on the ShapeNet test set using two NVIDIA RTX 2080

Ti GPU devices for different numbers of views. The results show that as the number of views increases, the inference time also increases due to the larger amount of data to be processed. However, WGA-SWIN consistently exhibits superior efficiency, outperforming both LRGT and UMIFormer across all view configurations.

Table 7. The average inference time (s) comparison between our WGA-SWIN and other previous methods, such as UMIFormer and LRGT, on the ShapeNet test set. To handle the variables, the same decoder as WGA-SWIN is used in all experiments.

Models	3 Views	8 Views	12 Views	16 Views	20 Views
UMIFormer [13]	0.0328 (s)	0.0494 (s)	0.0652 (s)	0.0818 (s)	0.0995 (s)
LRGT [11]	0.0319 (s)	0.0480 (s)	0.0639 (s)	0.0796 (s)	0.0983 (s)
WGA-SWIN (Ours)	**0.0302 (s)**	**0.0469 (s)**	**0.0627 (s)**	**0.0784 (s)**	**0.0973 (s)**

The best results are highlighted in bold.

For example, when processing three views, WGA-SWIN takes only 0.0302 (s), which is slightly faster than LRGT with 0.0319 (s) and UMIFormer with 0.0328 (s). This trend continues as the number of views increases, with WGA-SWIN maintaining the lowest inference time of 0.0973 (s) for 20 views, which is only slightly lower than the other two methods. Specifically, WGA-SWIN strikes an impressive balance between inference time and model performance, outperforming both LRGT and UMIFormer in terms of performance while maintaining similar inference times.

4.6. Failure Cases

The proposed model delivers impressive performance in static environments, but it fails to generate very accurate shapes in some scenarios. One key failure case occurs in dynamic scenes, where objects are in motion, or the environment undergoes rapid changes. Since WGA-SWIN is designed for static scenes, its reliance on fixed window-based attention groupings limits its ability to track moving objects or capture temporal changes. As a result, the model may fail to track moving objects accurately and handle interactions between objects across different frames, leading to sub-optimal performance in dynamic environments. Additionally, WGA-SWIN may fail to generate very accurate shapes in some cases of extreme occlusions or highly complex geometries. In scenarios where large portions of an object are hidden in all views and there are objects with intricate geometries or fine details, WGA's local window-based attention may fail to capture the missing information and smaller geometric details in complex objects. We expect to continue optimizing this scheme with learnable grouping techniques in future work to improve our model's robustness and performance in more complex real-world scenarios.

5. Conclusions

In this article, we design WGA-SWIN, a novel approach for multi-view 3D reconstruction that uses Window Grouping Attention (WGA) to efficiently capture intra- and inter-view dependencies using a swin transformer integrated into our encoder. In conjunction with a progressive hierarchical decoder that integrates swin transformer blocks with 3D convolutional layers, our method provides efficient voxel-level and high-resolution 3D reconstruction. According to the results of the experiments conducted on the benchmark ShapeNet dataset, the proposed method outperforms existing SOTA approaches in 3D reconstruction. We lead by 0.95% and 1.07% in both IoU and F-Scores, respectively, which demonstrates the effectiveness of our method. Our approach enhances multi-view feature aggregation and spatial consistency, making it ideal for high-resolution 3D object reconstruction tasks. In addition to providing an efficient and scalable solution with a variety of potential applications, multi-view 3D reconstruction is developed by WGA-SWIN. Even though our method simplifies computation using static grouping in WGA, it limits the flexibility of dynamically modeling relationships between views. Additionally, the method may struggle with extreme occlusions or highly detailed objects with complex geometries, where local window-based attention has limitations. Despite these challenges, WGA-SWIN provides a scalable and efficient solution for multi-view 3D reconstruction. We will investigate adaptive grouping techniques in the future to improve the performance of 3D object reconstruction on a variety of datasets and real-time applications.

Author Contributions: Conceptualization, S.S.M. and S.R.; methodology, S.S.M. and S.R.; software, S.S.M.; validation, S.S.M., M.Y.K.R. and S.R.; formal analysis, S.S.M. and S.R.; investigation, S.R.; resources, S.S.M. and S.R.; data curation, S.S.M. and G.F.A.; writing—original draft preparation, S.S.M.; writing—review and editing, S.R., M.Y.K.R. and G.F.A.; visualization, S.S.M., M.Y.K.R. and G.F.A.; supervision, S.R.; project administration, S.S.M. and S.R.; funding acquisition, S.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data will be made available upon request.

Acknowledgments: The authors would like to thank all individuals and institutions that contributed to the success of this work.

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

References

1. Wu, J.; Wyman, O.; Tang, Y.; Pasini, D.; Wang, W. Multi-view 3D reconstruction based on deep learning: A survey and comparison of methods. *Neurocomputing* **2024**, *582*, 127553. [CrossRef]
2. Özyeşil, O.; Voroninski, V.; Basri, R.; Singer, A. A survey of structure from motion. *Acta Numer.* **2017**, *26*, 305–364. [CrossRef]
3. Fuentes-Pacheco, J.; Ruiz-Ascencio, J.; Rendón-Mancha, J.M. Visual simultaneous localization and mapping: A survey. *Artif. Intell. Rev.* **2015**, *43*, 55–81. [CrossRef]
4. Kar, A.; Häne, C.; Malik, J. Learning a multi-view stereo machine. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 364–375.
5. Choy, C.B.; Xu, D.; Gwak, J.; Chen, K.; Savarese, S. 3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction. In *Computer Vision—ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 628–644.
6. Yang, B.; Wang, S.; Markham, A.; Trigoni, N. Robust Attentional Aggregation of Deep Feature Sets for Multi-view 3D Reconstruction. *Int. J. Comput. Vis.* **2020**, *128*, 53–73. [CrossRef]
7. Xie, H.; Yao, H.; Sun, X.; Zhou, S.; Zhang, S. Pix2Vox: Context-Aware 3D Reconstruction from Single and Multi-View Images. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2690–2698.
8. Wang, M.; Wang, L.; Fang, Y. 3DensiNet: A Robust Neural Network Architecture towards 3D Volumetric Object Prediction from 2D Image. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 961–969.
9. Xie, H.; Yao, H.; Zhang, S.; Zhou, S.; Sun, W. Pix2Vox++: Multi-scale Context-aware 3D Object Reconstruction from Single and Multiple Images. *Int. J. Comput. Vis.* **2020**, *128*, 2919–2935. [CrossRef]
10. Zhu, Z.; Yang, L.; Lin, X.; Yang, L.; Liang, Y. GARNet: Global-aware multi-view 3D reconstruction network and the cost-performance tradeoff. *Pattern Recognit.* **2023**, *142*, 109674. [CrossRef]
11. Yang, L.; Zhu, Z.; Nong, X.L.J.; Liang, Y. Long-Range Grouping Transformer for Multi-View 3D Reconstruction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023.
12. Wang, D.; Cui, X.; Chen, X.; Zou, Z.; Shi, T.; Salcudean, S.; Wang, Z.J.; Ward, R. Multi-view 3D Reconstruction with Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021.

13. Zhu, Z.; Yang, L.; Li, N.; Jiang, C.; Liang, Y. UMIFormer: Mining the Correlations between Similar Tokens for Multi-View 3D Reconstruction. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023.
14. Shi, Z.; Meng, Z.; Xing, Y.; Ma, Y.; Wattenhofer, R. 3D-RETR: End-to-End Single and Multi-View 3D Reconstruction with Transformers. In Proceedings of the 32nd British Machine Vision Conference, BMVC 2021, Online, 22–25 November 2021.
15. Yagubbayli, F.; Tonioni, A.; Tombari, F. LegoFormer: Transformers for Block-by-Block Multi-view 3D Reconstruction. *arXiv* **2021**, arXiv:2106.12102.
16. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16× 16 words: Transformers for image recognition at scale. In Proceedings of the ICLR 2021—9th International Conference on Learning Representations, Vienna, Austria, 3–7 May 2021.
17. Zhu, M.; Zhao, Z.; Cai, W. Hybrid Focal and Full-Range Attention Based Graph Transformers. In Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN), Yokohama, Japan, 30 June–5 July 2024; pp. 1–9.
18. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021.
19. Li, C.; Xiao, M.; Chen, F.; Qiao, S.; Wang, D.; Gao, M.; Zhang, S. R3D-SWIN:Use Shifted Window Attention for Single-View 3D Reconstruction. *arXiv* **2023**, arXiv:2312.02725.
20. Hossain, M.d.B.; Shinde, R.K.; Imtiaz, S.M.; Hossain, F.M.F.; Jeon, S.-H.; Kwon, K.-C.; Kim, N. Swin Transformer and the Unet Architecture to Correct Motion Artifacts in Magnetic Resonance Image Reconstruction. *Int. J. Biomed. Imaging.* **2024**, *2024*, 8972980. [CrossRef] [PubMed]
21. Kerbl, B.; Kopanas, G.; Leimkuehler, T.; Drettakis, G. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.* **2023**, *42*, 1–14. [CrossRef]
22. Hedman, P.; Srinivasan, P.P.; Mildenhall, B.; Barron, J.T.; Debevec, P. Baking Neural Radiance Fields for Real-Time View Synthesis. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021.
23. Ning, X.; Jiang, L.; Li, W.; Yu, Z.; Xie, J.; Li, L.; Tiwari, P.; Alonso-Fernandez, F. Swin-MGNet: Swin Transformer based Multi-view Grouping Network for 3D Object Recognition. *IEEE Trans. Artif. Intell.* **2024**, *6*, 747–758. [CrossRef]
24. Zeng, Y.; Fu, J.; Chao, H. Learning Joint Spatial-Temporal Transformations for Video Inpainting. In *Computer Vision—ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 528–543.
25. Liu, R.; Deng, H.; Huang, Y.; Shi, X.; Lu, L.; Sun, W.; Wang, X.; Dai, J.; Li, H. Decoupled Spatial-Temporal Transformer for Video Inpainting. *arXiv* **2021**, arXiv:2104.06637.
26. Li, G.; Cui, Z.; Li, M.; Han, Y.; Li, T. Multi-attention fusion transformer for single-image super-resolution. *Sci. Rep.* **2024**, *14*, 10222. [CrossRef] [PubMed]
27. Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; Guo, B. CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
28. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. SwinIR: Image Restoration Using Swin Transformer. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 1833–1844.
29. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
30. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
31. Chang, A.X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. ShapeNet: An Information-Rich 3D Model Repository. *arXiv* **2015**, arXiv:1512.03012.
32. Sun, X.; Wu, J.; Zhang, X.; Zhang, Z.; Zhang, C.; Xue, T.; Tenenbaum, J.B.; Freeman, W.T. Pix3D: Dataset and Methods for Single-Image 3D Shape Modeling. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
33. Tatarchenko, M.; Richter, S.R.; Ranftl, R.; Li, Z.; Koltun, V.; Brox, T. What do single-view 3D reconstruction networks learn? In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
34. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. *arXiv* **2017**, arXiv:1711.05101.

35. Su, H.; Qi, C.R.; Li, Y.; Guibas, L.J. Render for CNN: Viewpoint Estimation in Images Using CNNs Trained with Rendered 3D Model Views ILSVRC Image Classification Top-5 Error (%). In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
36. Xiao, J.; Hays, J.; Ehinger, K.A.; Oliva, A.; Torralba, A. SUN database: Large-scale scene recognition from abbey to zoo. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3485–3492.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.