applied sciences

MDPI

Article

SS3DNet-AF: A Single-Stage, Single-View 3D Reconstruction Network with Attention-Based Fusion

Muhammad Awais Shoukat [1], Allah Bux Sargano [1], Alexander Malyshev [2], Lihua You [3] and Zulfiqar Habib [1,*]

[1] Department of Computer Science, COMSATS University Islamabad, Lahore Campus, Lahore 54000, Pakistan; sp19-pcs-001@cuilahore.edu.pk (M.A.S.); allahbux@cuilahore.edu.pk (A.B.S.)
[2] Department of Mathematics, University of Bergen, 5020 Bergen, Norway; alexander.malyshev@uib.no
[3] National Centre for Computer Animation, Bournemouth University, Poole BH12 5BB, UK; lyou@bournemouth.ac.uk
* Correspondence: drzhabib@cuilahore.edu.pk

Abstract: Learning object shapes from a single image is challenging due to variations in scene content, geometric structures, and environmental factors, which create significant disparities between 2D image features and their corresponding 3D representations, hindering the effective training of deep learning models. Existing learning-based approaches can be divided into two-stage and single-stage methods, each with limitations. Two-stage methods often rely on generating intermediate proposals by searching for similar structures across the entire dataset, a process that is computationally expensive due to the large search space and high-dimensional feature-matching requirements, further limiting flexibility to predefined object categories. In contrast, single-stage methods directly reconstruct 3D shapes from images without intermediate steps, but they struggle to capture complex object geometries due to high feature loss between image features and 3D shapes and limit their ability to represent intricate details. To address these challenges, this paper introduces SS3DNet-AF, a single-stage, single-view 3D reconstruction network with an attention-based fusion (AF) mechanism to enhance focus on relevant image features, effectively capturing geometric details and generalizing across diverse object categories. The proposed method is quantitatively evaluated using the ShapeNet dataset, demonstrating its effectiveness in achieving accurate 3D reconstructions while overcoming the computational challenges associated with traditional approaches.

Keywords: SS3DNet-AF; 3D reconstruction; attention-based fusion; point clouds

Citation: Shoukat, M.A.; Sargano, A.B.; Malyshev, A.; You, L.; Habib, Z. SS3DNet-AF: A Single-Stage, Single-View 3D Reconstruction Network with Attention-Based Fusion. *Appl. Sci.* **2024**, *14*, 11424. https://doi.org/10.3390/app142311424

Academic Editor: Andrea Prati

Received: 27 October 2024
Revised: 27 November 2024
Accepted: 5 December 2024
Published: 8 December 2024

1. Introduction

Image-based 3D reconstruction has become essential across various industries, providing a cost-effective alternative to expensive equipment such as laser scanners and depth sensors. While 3D scanners and RGB-D cameras (depth sensors) are capable of capturing detailed 3D information, they come with significant limitations. These devices are often expensive, large in size, and require controlled environments or specific setups for accurate results. Additionally, they might not be practical for large-scale deployment, outdoor scenarios, or applications with strict budget constraints. Alternatively, simple image-capturing devices, such as standard cameras, are widely accessible and affordable, inspiring researchers to develop methods for reconstructing 3D shapes directly from images captured using these devices. While humans can naturally perceive the shape and structure of objects, replicating this ability in machines remains a significant research challenge. The applications of 3D reconstruction span diverse fields, including 3D character modeling, autonomous navigation for vehicles and robots for interacting with their surroundings, image-assisted surgical procedures where precise body scans are required, and 3D visualization of buildings and maps [1–5]. Despite the importance of 3D models, most imaging devices are still limited to capturing only two-dimensional (2D) information (x

and y coordinates), lacking the depth (*z*-axis) information essential for creating accurate 3D representations. Typically, 3D models are represented as voxel grids, point clouds, or meshes. However, manually creating these models is time-consuming and expensive, which promotes researchers to develop automated tools for generating 3D models directly from 2D images. Figure 1 illustrates a 2D image and its corresponding 3D model from various angles.

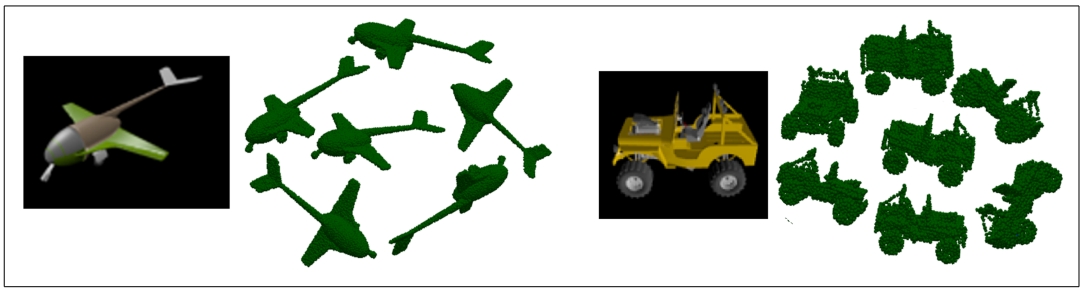

Figure 1. Examples from the ShapeNet dataset: 2D images of an airplane and a vehicle alongside their corresponding 3D models viewed from multiple angles.

When a camera captures an image, depth information is lost due to the projection of a 3D scene onto a 2D plane [6]. To address this issue, several methods have been developed to recover the missing dimension to accurately reconstruct 3D models—a critical task, as no direct cues about the lost dimension are available in the image plane. Broadly, 3D reconstruction methods that create 3D models directly from 2D images can be classified into conventional and learning-based methods. These methods are further categorized based on their applications, ranging from general object reconstruction to specialized domains like 3D facial modeling for gesture recognition, 3D human body reconstruction, character modeling, and medical imaging (see Figure 2). Key studies provide valuable insights into each of these research areas.

Traditional multi-view geometry methods, such as stereo vision and structure-from-motion, depend on capturing multiple images from different angles [7,8], a process that is both time-intensive and impractical for large-scale applications [1]. These methods also require significant expertise and extensive object coverage, limiting their accessibility. Alternatively, single-view learning-based approaches are categorized into two branches. The first involves generating intermediate 3D proposals by searching for structurally similar objects from extensive datasets, followed by fusion and refinement based on coarse estimates. Although this approach is effective, it is computationally expensive due to high-dimensional feature matching and a large search space, which makes it unsuitable for real-time applications. Additionally, it restricts models to predefined object categories, reducing adaptability. The second approach learns 3D shapes directly from image features, without involving intermediate steps and initial proposals, but it often struggles to capture intricate geometries due to significant feature disparities between images and their corresponding 3D representations.

To overcome these limitations, this paper introduces SS3DNet-AF, a novel single-stage, single-view 3D reconstruction network. Unlike traditional methods, SS3DNet-AF eliminates intermediate proposals by employing an attention-based fusion (AF) mechanism that seamlessly integrates image features with geometric information. The AF layers enhance the network's focus on the most relevant aspects of input images, enabling accurate 3D reconstructions across diverse object categories. The proposed SS3DNet-AF method was evaluated on a subset of the ShapeNet dataset, comprising 13 object categories of synthetic 3D models. Experimental results demonstrate the network's competitive accuracy and efficiency in addressing the computational limitations of traditional approaches. However, the proposed single-stage design encounters challenges in handling variations in lighting conditions, such as underexposure and overexposure, where object details become unclear, as well as occluded objects. Nevertheless, these limitations are less critical for most 3D

applications, where objects are typically fully visible, making the method suitable for a wide range of practical scenarios. Key contributions of this work include the following:

1. An improved single-stage framework for 3D reconstruction, eliminating the need for intermediate proposal generation and reducing computational complexity while maintaining competitive performance.
2. The development of an attention-based fusion (AF) mechanism that effectively integrates image features and geometric information, enhancing the reconstruction of 3D models from single-view images.

The remainder of this paper is structured as follows: Section 2 reviews related work, summarizing key advancements and challenges. Section 3 presents the proposed methodology in detail. Section 4, Experimentation and Results, covers the dataset overview, experimental setup, an ablation study examining the impact of the initial geometry representation, and an analysis of the attention mechanism's effectiveness by comparing single-stage and two-stage fusion layers. Finally, Section 5 concludes the study.

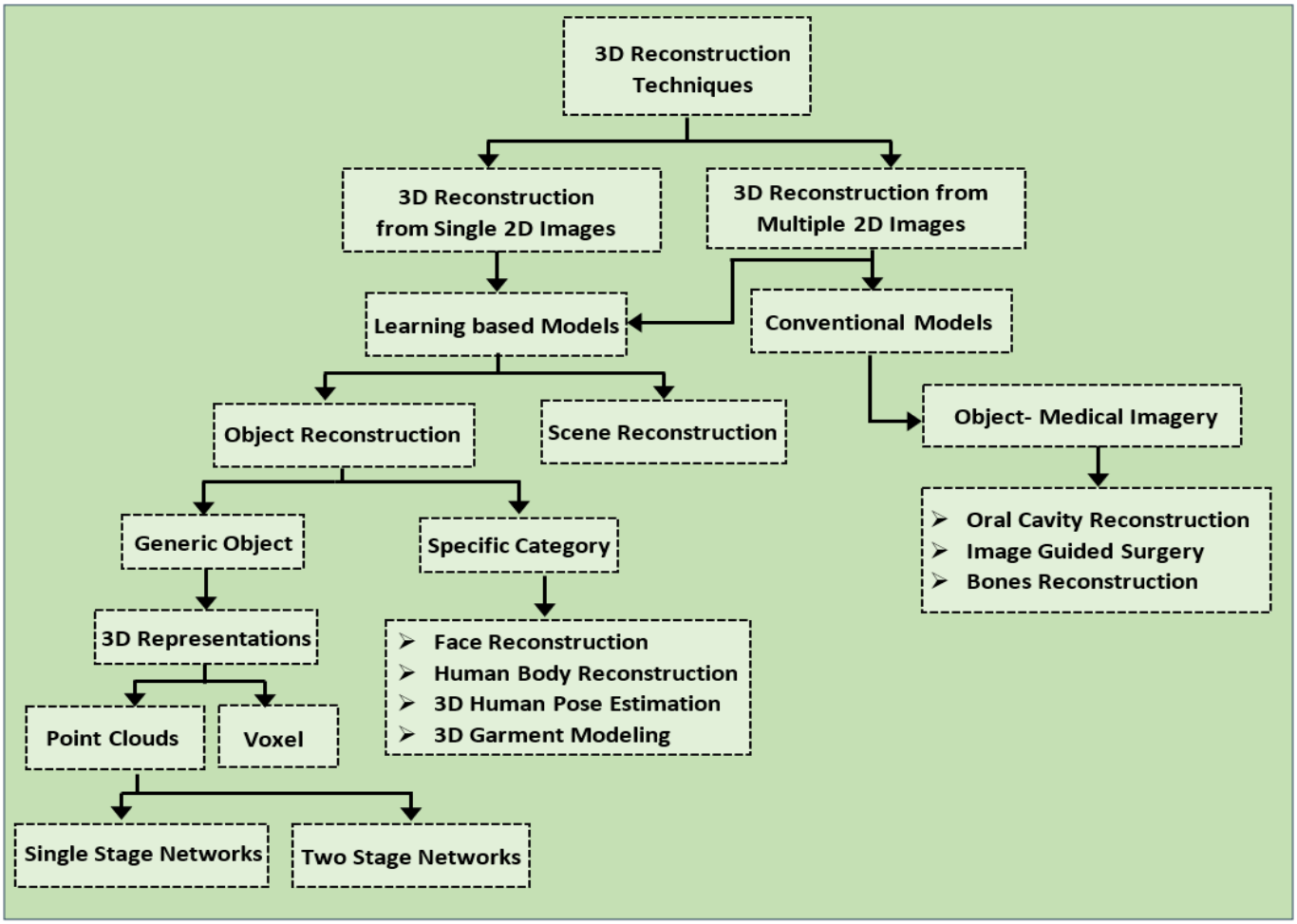

Figure 2. Taxonomy of 3D reconstruction techniques, categorized into methods based on single and multiple 2D images. Single-image approaches rely on learning-based models for object and scene reconstruction, utilizing 3D representations such as point clouds and voxels. In contrast, multi-image methods are based on conventional models, with notable applications in medical imaging, including oral cavity reconstruction, image-guided surgery, and bone reconstruction, etc.

2. Related Work

This section reviews 3D reconstruction techniques and how they have evolved from traditional geometry-based methods to advanced learning-based and hybrid approaches,

encouraging researchers to explore their potential for industrial applications, as discussed in the following subsections.

2.1. Conventional Techniques

Conventional 3D reconstruction methods rely on geometric principles like triangulation and epipolar geometry to estimate depth from multiple images [7,8]. The triangulation process estimates the depth of a point in space by calculating the intersection of the lines projected from different camera views. This process needs to know the relative location/pose of the cameras. Epipolar geometry makes it easy for this case because a point in one image corresponds to a line (the epipolar line) in the other image. Although these techniques yield accurate results in ideal conditions, they are computationally intensive and require thorough camera setup calibration [1].

2.2. Learning-Based Methods

With the advent of deep learning, there have been alternative methods aiming to estimate the missing depth cues directly from 2D images and reconstruct the 3D structures. Early methods, such as dual-stream neural networks [9], utilized coarse and refined streams to estimate depth maps. However, these methods commonly produced blurry depth maps as they used L2 loss during training. To improve performance, multi-view depth map fusion techniques were developed, projecting depth maps into unified 3D spaces [10]. Although effective, these methods added computational overhead by requiring intermediate depth map and multiple loss functions. Recent advancements focus on reconstructing 3D structures directly from single 2D images, addressing challenges like missing depth cues and the intrinsic complexity of geometric data [11–16]. Various approaches treat 3D reconstruction as a classification problem, employing diverse datasets to infer 3D shapes and structures. For example, some approaches assemble composite models by retrieving and deforming shape components [11,17]. However, extracting accurate 3D shapes directly from images is difficult due to wide variations in object geometries and structures.

Alternatively, some techniques adopt a modular approach, where segmentation masks, depth cues, and normal maps are treated as intermediate representations and then combined to reconstruct the object's geometry through geometric transformations (e.g., 3D back-projection) [1]. End-to-end architectures have further simplified the process by integrating intermediate stages [18,19], often using autoencoder–decoder frameworks to generate volumetric grids. Moreover, some methods incorporated shape priors to improve reconstruction accuracy [20,21]. The transition to end-to-end methods has reduced reliance on intermediate estimates, thereby minimizing accuracy loss. However, challenges persist, including memory limitations that restrict grid resolution to 1024^3. Recently, methods utilizing learnable 3D representations, such as voxels and point clouds, have achieved higher accuracy and enabled a broader range of applications. These advancements are explored in detail in the following subsections.

2.2.1. Voxel-Based Reconstruction

Voxel-based 3D reconstruction decomposes the 3D space into a volumetric grid of voxels that are considered occupied or empty and thus can be represented in a format suitable for deep learning frameworks. However, there are memory limitations that usually restrict the grid resolution to 1024^3 or less. The earliest works, like 3D-R2N2, used 3D convolutional networks paired with Long Short-Term Memory (LSTM) layers to output voxel-based representations from single or multiple images [12]. This approach uses an encoder, a 3D conv. LSTM, and a decoder to deal with multiple views of an object. Despite its innovative design, the model faces challenges; i.e., it has low resolution and a slow inference speed, which is mainly due to the sequential processing of images. To address these limitations, the OCTree Generating Network (OGN) introduced an adaptive voxel subdivision technique, which dynamically refines voxel grids to optimize memory usage, allowing for more detailed representations of 3D objects [22]. Further optimizations achieved

higher Intersection over Union (IoU) scores, even with fewer model parameters compared to previous methods, but as a result of the reduced number of model parameters, the resolution remains limited to 32^3 voxels [23]. Pix2Vox++ further enhanced voxel-based reconstruction by incorporating a ResNet backbone, which not only reduced parameters by 25% but also improved accuracy by 1.5% [24]. Additionally, variational autoencoders (VAEs) have played a significant role in generating smoother, higher-resolution 3D models. By learning latent feature representations from images, VAEs enable the reconstruction of more realistic 3D structures [25].

Recent approaches have further improved the accuracy and quality of 3D reconstructions. RNN-based discriminative networks, for example, improve the processing of sequential data, which is crucial for multi-view reconstruction [26], Generative Adversarial Networks (GANs) [27] have contributed to creating more detailed voxel grids through adversarial training, thus improving models' abilities to produce realistic 3D outputs, multi-scale context-aware fusion [23] enables models to capture both local and global features of objects, and transformers [28] provide robust capabilities in handling complex relationships between different parts of an object. Memory-based frameworks have also evolved to manage occlusions—situations where parts of an object are not visible in any provided view. These frameworks retain the memory of previously seen object parts, which helps to infer the structure of occluded areas [4]. Despite these advancements, voxel-based approaches continue to face challenges, such as low spatial resolution and high computational demands due to sparse voxel grids. These limitations continue to restrict their application in fields that require highly detailed and precise 3D reconstructions, such as medical imaging and virtual reality.

2.2.2. Point Cloud-Based Reconstruction

In parallel, significant progress has been made in point cloud generation, offering a memory-efficient way to represent 3D structures through an unordered set of points (x, y, z coordinates). Unlike structured representations, i.e., meshes that store connectivity information between vertices and edges, point clouds capture geometric details without the need for maintaining connectivity, which simplifies the process while preserving essential shape information. Early approaches in this domain utilized deep encoder–decoder architectures to estimate 3D point clouds from single images using Chamfer and Earth Mover's distances as loss functions to assess geometric similarity and shape accuracy [14]. Building on this, 3D-LMNet introduced a method for knowledge transfer between 2D and 3D domains, particularly targeting single-view reconstruction tasks. It uses a point cloud autoencoder trained with Chamfer distance (CD) loss to map 2D images into a learned 3D latent space. An extension with variational autoencoder (VAE) further allowed for generating multiple plausible outputs by capturing uncertainties inherent in single-image reconstructions [29]. Apart from GANs, other conditional flow-based networks have also been studied to achieve precise computation and flexible manipulation in the latent space [30]. Another approach, DensePCR [31], adopts a strategy to hierarchically refine low-resolution predictions through the aggregation of local and global features.

The single-encoder multiple-decoder (SE-MD) network [32] further advanced these techniques by employing a single autoencoder to learn feature distributions, coupled with multiple decoders to generate point clouds. These outputs were then fused to create detailed 3D reconstructions. It is similar to networks that utilize silhouettes as intermediate representations [33,34], which are combined with image features to generate point clouds for handling occluded objects [34]. More recently in this domain, both single-stage and two-stage networks have improved the accuracy of point cloud generation [35–38]. For instance, Pixel2point [37] introduced a single-stage network that uses an initial sphere point cloud to efficiently learn geometric shape parameters. Similarly, 3D-ReConstnet [35] applied a residual network and multilayer perceptions (MLPs) to extract features and predict point sets, using a learned Gaussian distribution to refine occluded regions. This method was later enhanced by using Detnet architecture with Exponential Linear Unit (ELU) activation, merging Earth Mover's distance (EMD) and Chamfer distance (CD) losses into a unified loss

function [38]. Subsequent advancements like 3D-CDRNet designed a two-stage point cloud reconstruction network that combines image features with a proposal retrieval branch. This model integrates an autoencoder, residual networks, and multilayer perceptrons (MLPs) to refine point set prediction, especially for occluded regions [36]. Although two-stage networks often outperform single-stage models, they tend to be computationally intensive and are limited to predefined categories. In contrast, single-stage networks are more efficient but often struggle to capture complex geometries due to the disparity between 2D image features and 3D representations. Considering their lightweight design and suitability for integration with IoT and real-time devices, our research focuses on improving single-stage networks.

However, unlike existing methods that either directly convert image features into 3D representations or merge them with initial point clouds, our approach strategically fuses spatial image features with initial spherical point cloud features. This fusion normalizes feature values, improving the learning process by bridging the gap between images and their corresponding 3D models. Additionally, SS3DNet-AF incorporates attention mechanisms to focus on relevant features while filtering out noise, enabling accurate capture of complex geometries and shapes. A detailed explanation of our methodology and technical specifications is provided in the following section.

3. Proposed Methods

This research proposes an automated system for generating 3D models directly from 2D images. Given an input image I, which may contain objects of varying categories, shapes, and geometries, our method learns a function f to estimate a 3D model M. This is achieved by intelligently fusing the features of I with an initial spherical geometry P, approximating the unknown model's ground truth Y. We use point clouds to represent the output (Equation (1)) and generate $N = 5000$ points from the CNN, which sufficiently captures the object's geometry and preserves its main structures. The process is formulated as follows:

$$\begin{aligned} M &= \#(P, I), \\ f &= loss(M, Y), \end{aligned} \tag{1}$$

where the hash symbol denotes the fusion mechanism discussed in module 3, Section 3.3. The loss function recognizes the generated and actual models with the help of Earth Mover's distance and Chamfer distance. This methodology is implemented through a four-module deep network, designed as follows:

(a) Image Encoder: Extracts high-level features from the input image.
(b) 3D Encoder: Processes the initial spherical point cloud to extract high-level 3D features while aligning their distribution with the image features.
(c) Dual Attention-Based Feature Fusion: Combines features from the image encoder and 3D encoder using a dual attention mechanism.
(d) Decoder: Generates the final point clouds from the fused features.

This structure ensures efficient and accurate 3D reconstruction. Below are the details of the network's modules as shown in Figure 3.

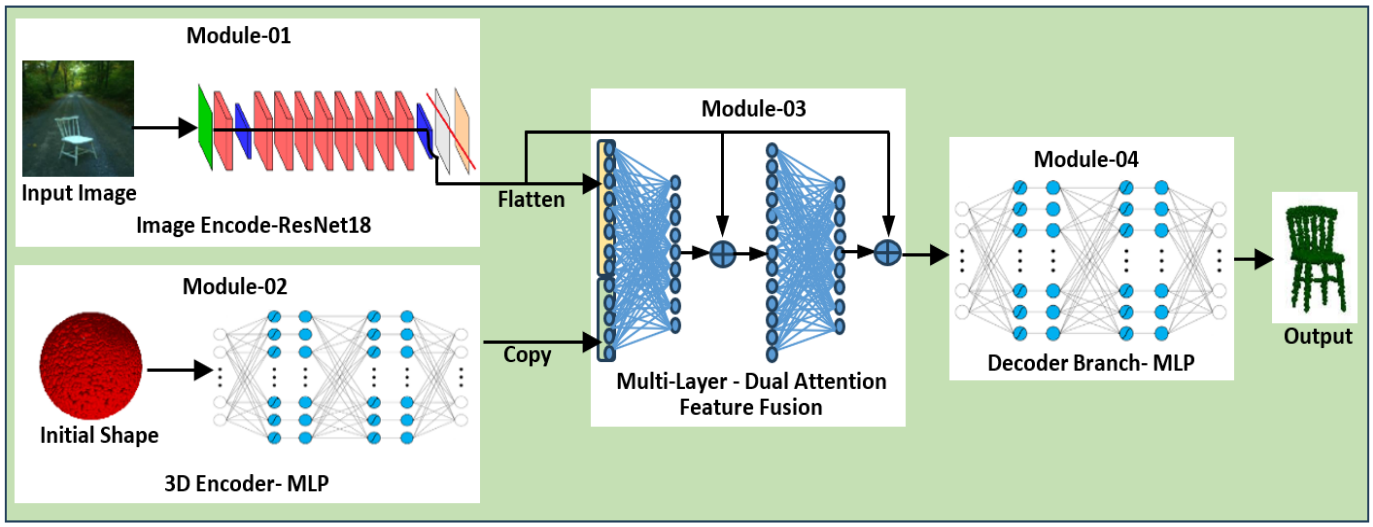

Figure 3. This Figure presents the proposed three-dimensional reconstruction network, SS3DNet-AF, designed to reconstruct a 3D shape from a 2D image. SS3DNet-AF consists of four modules. Module-01 employs ResNet18 to encode the 2D image, while Module-02 uses an MLP to encode the initial 3D shape (a spherical shape). In Module-03, the image and shape features are fused using a dual attention mechanism. Finally, Module-04 decodes the fused features with another MLP to produce the refined 3D shape.

3.1. Image Encoder

The image encoder module extracts high-level features from the input image using a pre-trained ResNet-18 model with a frozen classification layer. This module includes a convolutional layer, followed by four residual blocks, each having two convolutional layers and a global average pooling layer to flatten the extracted features. These residual blocks enhance the network's ability to capture and preserve key features. These features are then fed into the deep attention network, as detailed in Module-03. The network layers involved in this module are listed below.

3.1.1. Initial Convolution and Pooling

During the image encoding phase, we employ a convolutional operation and subsequent max pooling layer to analyze the input image I and extract essential low-level features. The mathematical representation of these operations is expressed as

$$\begin{aligned} X_{conv1} &= Conv(I, W_{conv1}), \\ X_{pool} &= MaxPool(X_{conv1}), \end{aligned} \tag{2}$$

where I is the input image, Conv signifies the convolution operation with a weight matrix W_{conv1}, and MaxPool denotes the max pooling function.

3.1.2. Residual Blocks

Four sequential residual blocks are utilized to learn and preserve significant high-level features:

$$\begin{aligned} X_{res1} &= ResBlock(X_{pool}, W_{res1}), \\ X_{res2} &= ResBlock(X_{res1}, W_{res2}), \\ X_{res3} &= ResBlock(X_{res2}, W_{res3}), \\ X_{res4} &= ResBlock(X_{res3}, W_{res4}), \end{aligned} \tag{3}$$

where, ResBlock denotes the operations within a residual block, and W_{res1}, W_{res2}, W_{res3}, and W_{res4} each represent block weights.

3.1.3. Global Average Pooling and Flattening

After the residual blocks, the global average pooling layer computes the average across all channels of the features, subsequently flattening them to create a feature vector. The operations are articulated as

$$\begin{aligned} X_{avgpool} &= AvgPool(X_{res4}), \\ X_I &= Flatten(X_{avgpool}), \end{aligned} \tag{4}$$

where *AvgPool* represents average pooling layer and *Flatten* represents the flattening of features.

3.2. 3D Encoder

The 3D encoder module learns the initial point cloud geometry using a multi-layer perceptron (MLP). It starts with a flattening layer, followed by three fully connected layers (4096, 1024, and 512 neurons) with the Leaky ReLU activation function having slope of 0.2, which introduces non-linearity to the network. The layers of the network are as follows:

Flatten Layer: This layer receives an initial sphere point cloud with dimensions (5000×3) as input and transforms it into a 15,000 × 1 feature vector by flattening the data.

$$X_{flat} = Flatten(P). \tag{5}$$

Linear Layers with Leaky Relu Activation: The feature vector from the previous layer is then passed through three multi-layer perceptron (MLP) layers to learn features and geometric parameters, as described by the following equations

$$\begin{aligned} X_{l1} &= X_{flat}.W_1 + b_1, \\ X_{l1} &= max(0.2 * (X_{l1}), X_{l1}), \\ X_{l2} &= X_{l1}.W_2 + b_2, \\ X_{l2} &= max(0.2 * (X_{l2}), X_{l2}), \\ X_P &= X_{l2}.W_3 + b_3, \end{aligned} \tag{6}$$

where W_1, W_2, and W_3 are the weight matrices with dimensions $(5000 \times 3) \times 4096$, (4096×1024), and (1024×512), and b_1, b_2, and b_3 are the biases on each layer, respectively.

3.3. Deep Attention-Based Feature Fusion

The deep attention-based feature fusion module combines features from two different sources: an image-encoding branch and a 3D-encoding branch, as discussed earlier. The image-encoding branch focuses on extracting features from the input image, while the 3D-encoding branch processes initial point cloud data to capture high-level geometric information. By merging these two types of data, the module creates a more complete and detailed understanding of the input.

The fusion process first merges the output of the two branches through a linear layer, reducing the combined dimensionality from 512 to 256. This fused representation is then merged with the output of the image-encoding branch through a second fusion layer, further refining the features into a 128-dimensional vector. Finally, the third fusion layer processes this refined representation to produce the final 512-dimensional fused features. Sequential fusion of these layers is critical for integrating information from both branches, resulting in a comprehensive feature representation. To enhance this process, attention mechanisms are embedded in each fusion layer. This approach helps the model focus on the most important features in the image-encoding branch, ensuring better generalization of the input image and varying inputs. The mathematical formulation of the fusion layers is as follows:

$$
\begin{aligned}
X_{f1} &= [X_I, X_P].W_{f1} + b_{f1}, \\
X_{f2} &= [X_I, X_{f1}].W_{f2} + b_{f2}, \\
X_{fused} &= [X_I, X_{f2}].W_{f3} + b_{f3},
\end{aligned} \tag{7}
$$

where X_I represents the features from the image-encoding branch (computed in Module-01), and X_P represents the features from the 3D-encoding branch (computed in Module-02). The weight matrices W_{f1}, W_{f2}, and W_{f3} are responsible for transforming and combining the feature representations at each stage. The dimensions of these matrices are as follows: W_{f1} has dimensions $(512 + 512) \times 256$, W_{f2} has dimensions $(512 + 256) \times 128$, and W_{f3} has dimensions $(512 + 128) \times 512$. Bias terms b_{f1}, b_{f2}, and b_{f3} are applied to each layer to fine-tune their outputs. This fusion method effectively fuses image features with shape parameters, resulting in a robust and unified feature representation that accurately captures the object's shape and geometry.

3.4. Decoder Branch

The decoder branch sequentially applies linear layers with Leaky ReLU activations following the fusion process, gradually increasing the feature dimensions from 512 to 1024, 8192, and finally to a 3D representation with dimensions of 5000×3. This architecture enhances the model's capability to capture complex geometric structures and helps it to learn highly detailed 3D models. The mathematical formulation of the decoder layers is provided in the following equations.

$$
\begin{aligned}
X_{D1} &= X_{fused}.W_{D1} + b_{D1}, \\
X_{D1} &= max(0.2 * (X_{D1}), X_{D1}), \\
X_{D2} &= X_{D1}.W_{D2} + b_{D2}, \\
X_{D2} &= max(0.2 * (X_{D2}), X_{D2}), \\
X_{D3} &= X_{D2}.W_{D3} + b_{D3}, \\
X_{D3} &= max(0.2 * (X_{D3}), X_{D3}), \\
M &= X_{D3}.W_{D4} + b_{D4},
\end{aligned} \tag{8}
$$

where M represents the model's output, the weight matrices W_{D1}, W_{D2}, W_{D3}, and W_{D4} have dimensions of (512×1024), (1024×4096), (4096×8192), and $(8192 \times 5000) \times 3$, respectively, with corresponding biases b_{D1}, b_{D2}, b_{D3}, and b_{D4} applied to each layer. These equations illustrate the forward pass through the neural network architecture. Details regarding the backward propagation process, gradient descent, and optimization parameters, such as learning rate, batch size, and regularization techniques, are discussed in Section 4.

4. Experimentation and Results

This section provides an overview of the dataset used in the experiment and details about the experimental setup, including parameter optimization and the use of gradient loss functions during network training.

4.1. Dataset Overview

To evaluate our method, we conducted experiments using the ShapeNet dataset [39], a well-known benchmark for 3D shape analysis in computer vision and imaging. The dataset contains 51,000 instances in 55 categories, such as furniture, vehicles, and animals. We selected ShapeNet due to its large size and organized structure, making it suited for detailed analysis. In this study, we focused on 13 categories (Airplane, Speaker, Cabinet, Monitor, Car, Chair, Rifle, Sofa, Table, Bench, Lamp, Vessel, and Telephone) and generated 2D views from 3D CAD models. These categories were chosen to make the fare comparison with the other SOFT methods. The dataset's clean backgrounds and consistent object views helped ensure effective training and reliable evaluation.

4.2. Experiment Setup

The experiments were performed on a DELL Precision-7910 machine with an NVIDIA GeForce GTX 1080 Ti GPU, 96 GB of RAM, and the PyTorch deep learning framework. The model architecture features two distinct branches: an image branch utilizing a pre-trained ResNet-18 for feature extraction with adaptive average pooling, flattening, and normalization steps, while the point cloud branch consists of fully connected layers with leaky ReLU activations to process 3D point cloud data. Features from both branches are fused through three linear transformation layers using attention mechanisms. Finally, the fused features are decoded using linear layers to generate the final 3D point cloud. During training, L2 normalization is applied to both image and point cloud features, and the model is optimized using stochastic gradient descent (SGD). Hyperparameters were fine-tuned throughout the experiments to minimize the loss function.

4.3. Ablation Study

To evaluate the contributions of key components in SS3DNet-AF, we conducted an ablation study using the ShapeNet dataset, with Chamfer distance (CD) as the primary evaluation metric. This study investigates two major aspects: the impact of the attention mechanism in feature fusion and the significance of the initial geometry representation.

4.3.1. Effectiveness of Attention Mechanism: Single-Stage, Two-Stage, and Multi-Level Fusion

The attention mechanism in the fusion module was evaluated by comparing a single-stage design to two-stage and multi-stage variations. The multi-stage approach shows improved reconstruction accuracy, and it was observed that removing the attention mechanism from the fusion module led to a significant decrease in performance, as the network struggled to capture details of the object effectively. Figures 4–6 present the 3D reconstruction results for various fusion and attention configurations. In Figure 4, the first column shows the input 2D images of different aircraft. The second column presents the results of direct feature fusion, which produces scattered and incomplete 3D shapes. The third column shows reconstructions using a single-layer attention mechanism, which moderately improves 3D structure but shows limitations in capturing shape details. The fourth column incorporates a two-layer attention mechanism, leading to better-defined shapes with enhanced spatial consistency. Lastly, the fifth column presents results from the proposed multi-level attention mechanism, achieving the most accurate and detailed reconstructions.

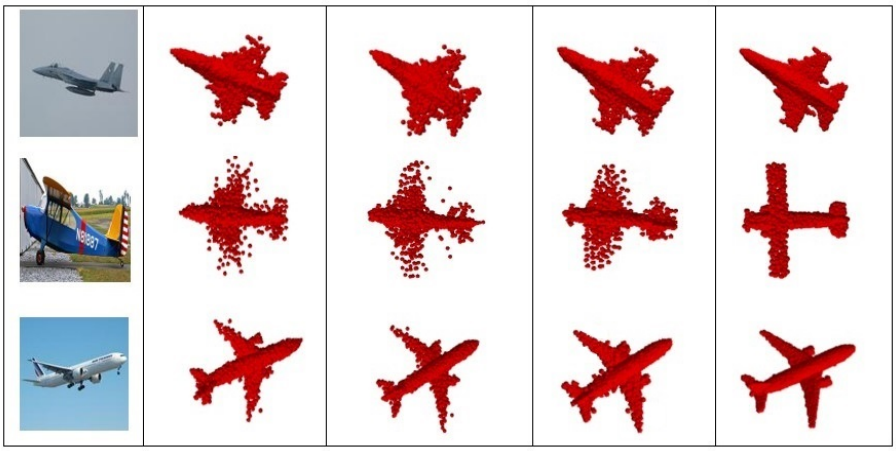

Figure 4. Comparison of 3D reconstruction results under various fusion and attention configurations. The input images (first column) are followed by reconstructions using direct feature fusion (second column, incomplete shapes), single-layer attention (third column, moderate improvement), two-layer attention (fourth column, enhanced spatial consistency), and multi-layer attention (fifth column, most accurate reconstruction).

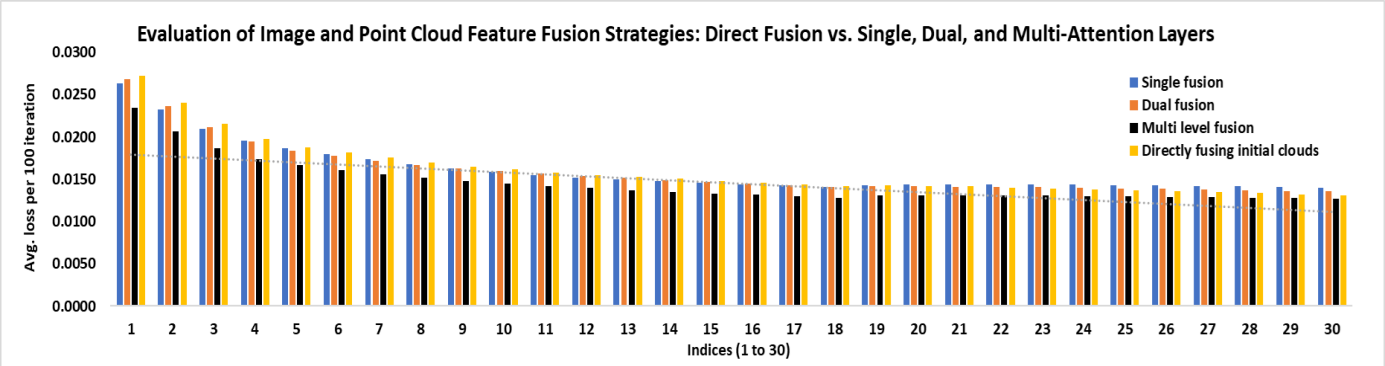

Figure 5. This Figure illustrates a comparison of average loss per 100 iterations across various attention-based feature fusion strategies, including single-attention fusion, dual-attention fusion, multi-level attention fusion, and direct fusion of initial clouds. Among these strategies, multi-level attention fusion achieves the lowest average loss, highlighting its effectiveness in minimizing reconstruction error. The dotted line represents the trend of the multi-level attention fusion model, indicating a continuous reduction in loss over the iterations. The *y*-axis represents the average loss per 100 iterations, while the *x*-axis corresponds to the value indices for the first 30 computed values.

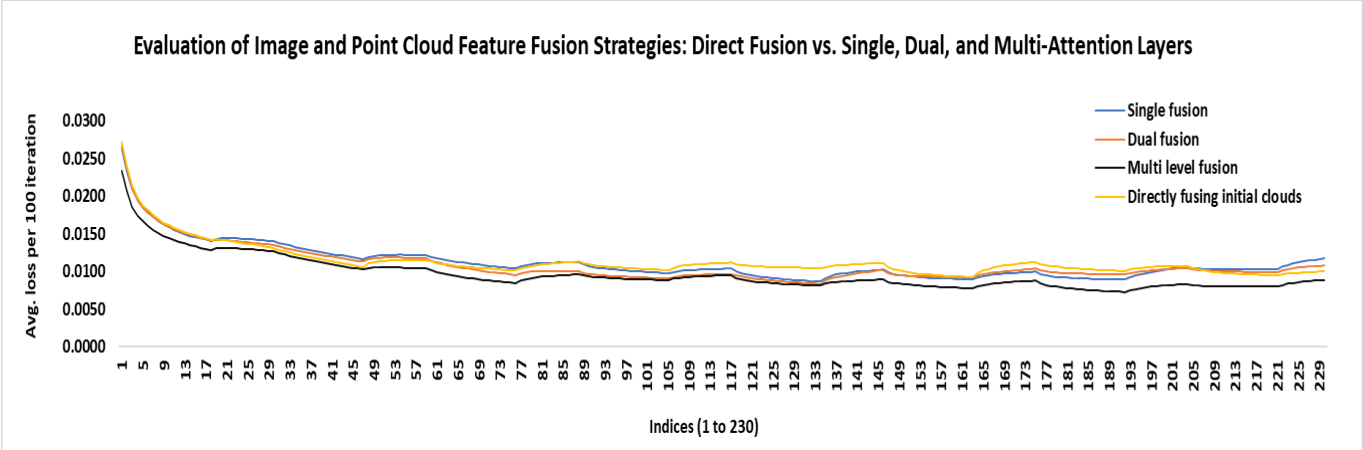

Figure 6. The line graph shows the average loss per 100 iterations for various attention-based feature fusion strategies. Multi-level attention fusion consistently achieves the lowest loss, outperforming dual-attention and single-attention fusion. In contrast, the direct fusion of initial clouds resulted in the highest loss. The *y*-axis represents the average loss per 100 iterations, while the *x*-axis denotes the value indices for the first 230 computed values.

4.3.2. Impact of Initial Geometry Representation

The effect of the initial geometry representation was examined by replacing the spherical point cloud, used in the proposed method, with a randomly distributed point cloud. This substitution caused a significant drop in performance, primarily due to the misalignment between image features and the random 3D grid, which resulted in erratic and fragmented reconstructions. The 3D reconstruction process under different geometric initializations is visualized in Figure 7. The first column represents the input 2D image, and the second column shows the results of the initial random point cloud geometry, which lacks meaningful structure and leads to poor-quality outputs. The third column shows results with spherical initial geometry, introducing some structural consistency, but it is still not enough for accurate reconstructions. Finally, the fourth column shows the integration of the spherical prior with attention-based fusion layers, leading to more precise and accurate 3D shapes. These results highlight the importance of combining prior geometric knowledge with attention mechanisms to achieve high-quality single-view 3D reconstructions.

Figure 7. Visualization of the 3D reconstruction process. The input image (first column) is followed by the initial random point cloud geometry (second column), the fusion with spherical prior for structural consistency (third column), and the final refinement using attention-based fusion layers (fourth column). This highlights the importance of geometric priors and attention mechanisms in achieving accurate and detailed reconstructions.

The detailed analyses validate the design choices in SS3DNet-AF and show how the dual attention mechanism and the induction of geometric prior enhance the quality of single-view 3D reconstructions. Furthermore, the model is evaluated on completely unseen data, as illustrated in Figure 8, which presents three sample objects (a car, a sofa, and a plane). The results present the model's ability to generalize effectively and produce accurate 3D reconstructions even for previously unseen inputs.

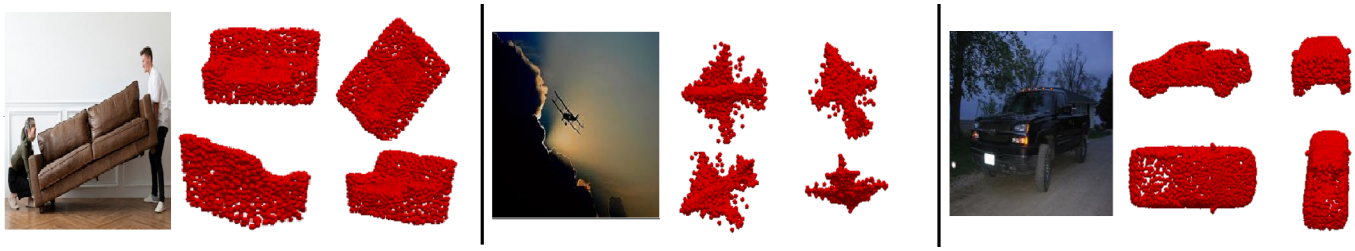

Figure 8. 3D point clouds generated from input images on previously unseen data. Each input image is paired with its corresponding 3D point cloud, visualized from multiple perspectives. These results highlight the model's capability to generalize and accurately reconstruct 3D shapes.

4.4. Parameters Optimization and Gradient Loss

Parameter optimization is a crucial step in updating model weights during training. The Chamfer distance (CD) was used as the primary loss function to evaluate the dissimilarity between the predicted point clouds M and the ground truth point clouds Y. CD computes the average of the minimum squared Euclidean distances between points in the two sets. For each point in M, the nearest point in Y is identified, and their squared distances are summed. This process is repeated for points in Y with respect to M. The final Chamfer distance, a standard metric for point cloud reconstruction, guides the model towards more accurate predictions.

$$\begin{aligned} loss &= \sum_{m \in M} \min_{y \in Y} \|m - y\|^2 + \sum_{y \in Y} \min_{m \in M} \|m - y\|^2, \\ \theta_{new} &= \theta_{old} - \alpha \cdot \frac{\partial_{loss}}{\partial_{\theta}}, \end{aligned} \tag{9}$$

where θ_{new} denotes the updated parameter, θ_{old} represents the current parameter, and α is the learning rate (set to 0.0001), which was tuned through various experiments. The results showed that a learning rate of 3×10^{-2} achieved 63.13%, 3×10^{-3} achieved 79.19%, and 3×10^{-4} achieved 81.09%, with the lowest learning rate yielding the highest accuracy, and $\nabla L(\theta_{old})$ indicates the gradient of the loss function. The Adam optimizer [40] was employed to update the parameters, providing a robust, adaptive approach that adjusts learning rates dynamically based on historical gradient information, enabling efficient convergence with minimal hyperparameter tuning. In this context, the integration of the

Adam optimizer with the Chamfer distance loss ensures stable and effective parameter optimization, ultimately enhancing 3D reconstruction performance.

Figure 9 displays the training loss on the dataset and Table 1 compares our method with state-of-the-art approaches like 3D-Reconstnet [35], 3D-FEGNET [41], DetNet [38], and Pixel2Point [37]. Results from other models were obtained from their publications or cited references. The proposed method shows better or more competitive results compared to most state-of-the-art algorithms, attributable to the deep fusion attention mechanism applied across both the image and point cloud branches. However, high reconstruction loss was observed for some object categories, which may be due to intricate geometries and fine details. Further analysis is needed to determine the precise reasons behind this performance gap. We will analyze these categories in our future work and will improve the results for these categories. Figure 10 presents the 3D point clouds generated by the proposed architecture.

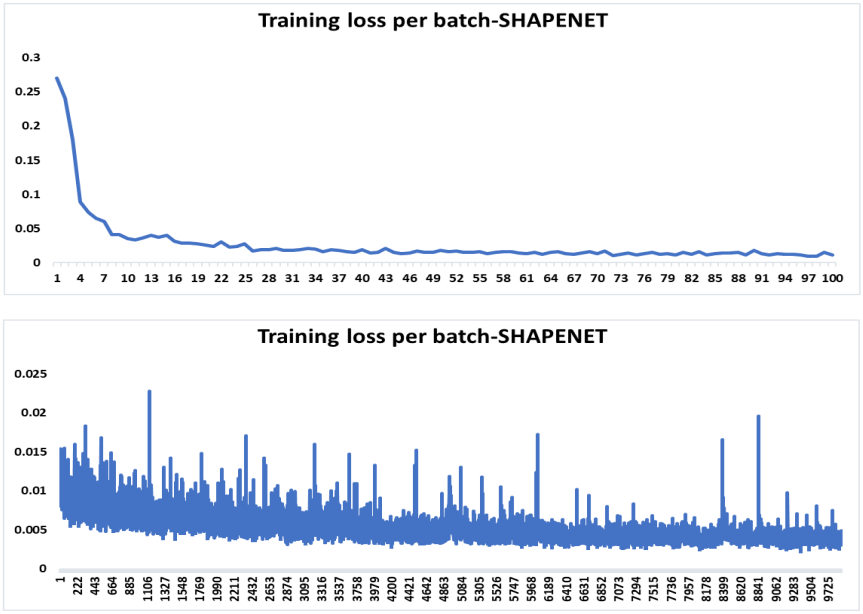

Figure 9. Training loss on the ShapeNet dataset: the first image displays the loss over the first 100 iterations, while the second image shows the loss for the remaining iterations.

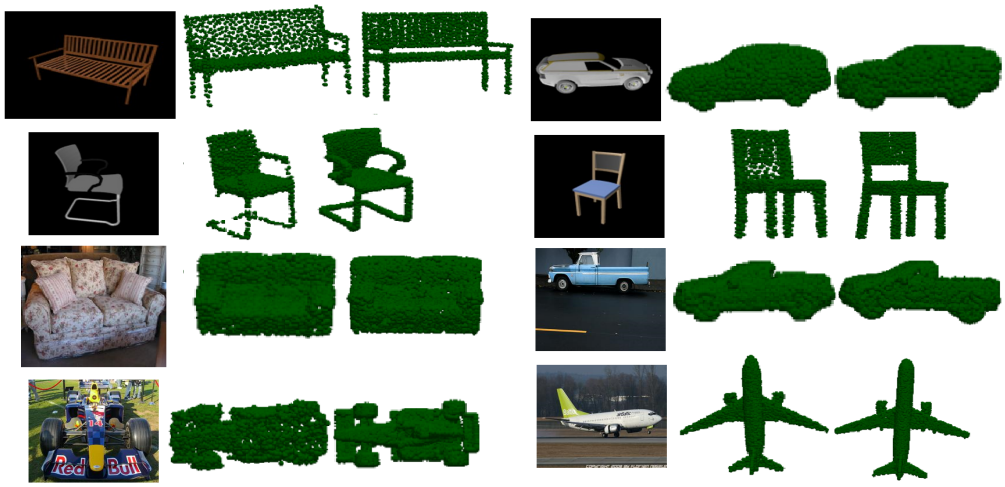

Figure 10. To assess model performance across diverse datasets with varying backgrounds, the Figure presents input images alongside the generated 3D models and their corresponding ground truth 3D representations.

Table 1. Comparison of Chamfer distance (CD) values for 3D models generated by our method and state-of-the-art techniques across ShapeNet object categories. Lower CD values indicate better performance. Methods compared include 3D-Reconstnet, 3D-FEGNET, DetNet, and Pixel2Point.

CD Comparison		Methods				
		3D-Reconstnet [35]	3D-FEGNET [41]	DetNet [38]	Pixel2Point [37]	Proposed
Category	Airplane	2.42	2.36	2.38	3.29	2.29
	Bench	3.57	3.60	3.51	4.59	3.50
	Cabinet	4.66	4.84	4.77	6.07	4.63
	Car	3.59	3.57	3.56	4.39	4.11
	Chair	4.41	4.35	4.35	6.48	4.32
	Lamp	5.03	5.13	4.99	6.58	4.96
	Monitor	4.61	4.67	4.72	6.39	4.58
	Rifle	2.51	2.45	2.45	2.89	2.61
	Sofa	4.58	4.56	4.44	5.85	4.32
	Speaker	5.94	6.00	5.94	8.39	5.91
	Table	4.41	4.42	4.35	6.26	4.28
	Telephone	3.59	3.50	3.52	4.27	3.46
	Vessel	3.81	3.75	3.72	4.55	3.76

5. Conclusions

This study presents SS3DNet-AF, a novel attention-based fusion network for 3D reconstruction from single-view images. By embedding attention mechanisms within the fusion layers, the network retains essential spatial information while enhancing feature fusion, leading to more accurate 3D reconstructions and effective handling of complex shapes. Experimental results on the ShapeNet dataset confirm that SS3DNet-AF outperforms several state-of-the-art models. Future research will focus on integrating part-level fusion to address deformations in unseen object components and will further investigate the impact of attention layers on other network branches to improve overall performance.

Author Contributions: M.A.S., A.B.S., A.M., L.Y. and Z.H. collaborated in conceptualizing and designing the research. M.A.S. led the methodology, executed the experiments, and prepared the manuscript. A.B.S. and Z.H. contributed by analyzing the data and refining the architecture. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

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

References

1. Han, X.F.; Laga, H.; Bennamoun, M. Image-based 3D object reconstruction: State-of-the-art and trends in the deep learning era. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1578–1604. [CrossRef] [PubMed]
2. Sra, M.; Garrido-Jurado, S.; Schmandt, C.; Maes, P. Procedurally generated virtual reality from 3D reconstructed physical space. In Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology, Munich, Germany, 2–4 November 2016; pp. 191–200.

3. Montefusco, L.B.; Lazzaro, D.; Papi, S.; Guerrini, C. A fast compressed sensing approach to 3D MR image reconstruction. *IEEE Trans. Med. Imaging* **2010**, *30*, 1064–1075. [CrossRef] [PubMed]
4. Yang, S.; Xu, M.; Xie, H.; Perry, S.; Xia, J. Single-view 3D object reconstruction from shape priors in memory. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3152–3161.
5. Pang, H.E.; Biljecki, F. 3D building reconstruction from single street view images using deep learning. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102859. [CrossRef]
6. Hartley, R.; Zisserman, A. *Multiple View Geometry in Computer Vision*; Cambridge University Press: Cambridge, UK, 2003; p. vi+560, ISBN 0-521-54051-8.
7. Schonberger, J.L.; Frahm, J.M. Structure-from-motion revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4104–4113.
8. Cadena, C.; Carlone, L.; Carrillo, H.; Latif, Y.; Scaramuzza, D.; Neira, J.; Reid, I.; Leonard, J.J. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Trans. Robot.* **2016**, *32*, 1309–1332. [CrossRef]
9. Eigen, D.; Puhrsch, C.; Fergus, R. Depth map prediction from a single image using a multi-scale deep network. *arXiv* **2014**, arXiv:1406.2283.
10. Tatarchenko, M.; Dosovitskiy, A.; Brox, T. Multi-view 3D models from single images with a convolutional network. In Proceedings of the European Conference on Computer Vision, Cham, Switzerland, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 322–337. [CrossRef]
11. Huang, Q.; Wang, H.; Koltun, V. Single-view reconstruction via joint analysis of image and shape collections. *ACM Trans. Graph. TOG* **2015**, *34*, 87–91. [CrossRef]
12. Choy, C.B.; Xu, D.; Gwak, J.; Savarese, S. 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 628–644. [CrossRef]
13. Girdhar, R.; Fouhey, D.F.; Rodriguez, M.; Gupta, A. Learning a predictable and generative vector representation for objects. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 484–499.
14. Fan, H.; Su, H.; Guibas, L.J. A point set generation network for 3D object reconstruction from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 605–613.
15. Feng, Y.; Wu, F.; Shao, X.; Wang, Y.; Zhou, X. Joint 3D face reconstruction and dense alignment with position map regression network. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 534–551.
16. Sinha, A.; Unmesh, A.; Huang, Q.; Ramani, K. Surfnet: Generating 3D shape surfaces using deep residual networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6040–6049.
17. Xiang, Y.; Kim, W.; Chen, W.; Ji, J.; Choy, C.; Su, H.; Mottaghi, R.; Guibas, L.; Savarese, S. Objectnet3D: A large scale database for 3D object recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 160–176.
18. 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 Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2974–2983.
19. Richter, S.R.; Roth, S. Matryoshka networks: Predicting 3D geometry via nested shape layers. In Proceedings of the IEEE Conference on Computer vision And Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1936–1944.
20. Wu, J.; Wang, Y.; Xue, T.; Sun, X.; Freeman, B.; Tenenbaum, J. Marrnet: 3D shape reconstruction via 2.5 D sketches. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
21. Wu, J.; Zhang, C.; Zhang, X.; Zhang, Z.; Freeman, W.T.; Tenenbaum, J.B. Learning shape priors for single-view 3D completion and reconstruction. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 646–662.
22. Tatarchenko, M.; Dosovitskiy, A.; Brox, T. Octree generating networks: Efficient convolutional architectures for high-resolution 3D outputs. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2088–2096.
23. Xie, H.; Yao, H.; Sun, X.; Zhou, S.; Zhang, Y. Pix2Vox: Context-aware 3D reconstruction from single and multi-view images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2690–2698. [CrossRef]
24. Xie, H.; Yao, H.; Zhang, S.; Zhou, S.; Sun, X. Pix2Vox++: Multi-scale context-aware 3D object reconstruction from single and multiple images. *Int. J. Comput. Vis.* **2020**, *128*, 2919–2935. [CrossRef]
25. Tahir, R.; Sargano, A.B.; Habib, Z. Voxel-based 3D object reconstruction from single 2D image using variational autoencoders. *Mathematics* **2021**, *9*, 2288. [CrossRef]
26. Han, Z.; Qiao, G.; Liu, Y.S.; Zwicker, M. SeqXY2SeqZ: Structure learning for 3D shapes by sequentially predicting 1D occupancy segments from 2D coordinates. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 607–625.

27. Kniaz, V.V.; Knyaz, V.A.; Remondino, F.; Bordodymov, A.; Moshkantsev, P. Image-to-voxel model translation for 3D scene reconstruction and segmentation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 105–124.
28. Peng, K.; Islam, R.; Quarles, J.; Desai, K. Tmvnet: Using transformers for multi-view voxel-based 3D reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 222–230.
29. Mandikal, P.; Navaneet, K.; Agarwal, M.; Babu, R.V. 3D-lmNET: Latent embedding matching for accurate and diverse 3D point cloud reconstruction from a single image. *arXiv* **2018**, arXiv:1807.07796.
30. Pumarola, A.; Popov, S.; Moreno-Noguer, F.; Ferrari, V. C-flow: Conditional generative flow models for images and 3D point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7949–7958.
31. Mandikal, P.; Radhakrishnan, V.B. Dense 3D point cloud reconstruction using a deep pyramid network. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 7–11 January 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1052–1060. [CrossRef]
32. Mueed Hafiz, A.; Alam Bhat, R.U.; Parah, S.A.; Hassaballah, M. SE-MD: A Single-encoder multiple-decoder deep network for point cloud generation from 2D images. *arXiv* **2021**, arXiv:2106.15325.
33. Laurentini, A. The visual hull concept for silhouette-based image understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **1994**, *16*, 150–162. [CrossRef]
34. Zou, C.; Hoiem, D. Silhouette guided point cloud reconstruction beyond occlusion. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Snowmass, CO, USA, 1–5 March 2020; pp. 41–50. [CrossRef]
35. Li, B.; Zhang, Y.; Zhao, B.; Shao, H. 3D-ReConstnet: A single-view 3D-object point cloud reconstruction network. *IEEE Access* **2020**, *8*, 83782–83790. [CrossRef]
36. Tong, Y.; Chen, H.; Yang, N.; Menhas, M.I.; Ahmad, B. 3D-CDRNet: Retrieval-based dense point cloud reconstruction from a single image under complex background. *Displays* **2023**, *78*, 102438. [CrossRef]
37. Afifi, A.J.; Magnusson, J.; Soomro, T.A.; Hellwich, O. Pixel2Point: 3D object reconstruction from a single image using CNN and initial sphere. *IEEE Access* **2020**, *9*, 110–121. [CrossRef]
38. Li, B.; Zhu, S.; Lu, Y. A single stage and single view 3D point cloud reconstruction network based on DetNet. *Sensors* **2022**, *22*, 8235. [CrossRef] [PubMed]
39. 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.
40. Kingma, D.P. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
41. Wang, E.; Sun, H.; Wang, B.; Cao, Z.; Liu, Z. 3D-FEGNet: A feature enhanced point cloud generation network from a single image. *IET Comput. Vis.* **2023**, *17*, 98–110. [CrossRef]