electronics

MDPI

Article

DepthCloud2Point: Depth Maps and Initial Point for 3D Point Cloud Reconstruction from a Single Image

Galana Fekadu Asafa, Shengbing Ren *, Sheikh Sohan Mamun and Kaleb Amsalu Gobena

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

Abstract: Reconstructing 3D objects from single-view images has acquired significant interest due to their wide-ranging applications in robotics, autonomous vehicles, virtual reality, and augmented reality. Current methods, including voxel-based and point cloud-based approaches, face critical challenges such as irregular point distributions and an inability to preserve complex object details, which result in suboptimal reconstructions. To address these limitations, we propose DepthCloud2Point, a framework that combines depth maps, image features, and an initial point cloud to generate detailed and accurate 3D point clouds. Depth maps are employed to provide rich spatial cues, resolving depth ambiguities, while the initial point cloud serves as a geometric prior to ensure uniform point distribution. These components are integrated into a unified pipeline, where the encoder extracts semantic and geometric features, and the generator synthesizes high-fidelity 3D reconstructions. Our approach is trained end-to-end on both synthetic and real-world datasets, achieving state-of-the-art performance. Quantitative results on the ShapeNet dataset show that DepthCloud2Point outperforms 3D-LMNet by 19.07% in CD and 38.86% in EMD, and Pixel2Point by 18.77% in CD and 19.25% in EMD. We also perform a qualitative study that shows that our approach is able to generate reconstructions that closely align with ground truth, capturing intricate object details and maintaining spatial coherence, confirming its outperforming over the 3D-LMNet and Pixel2Point.

Keywords: 3D reconstruction; DepthCloud2Point; depth maps; point cloud

Academic Editors: Di Wang, Huan Deng and Yilong Li

Received: 4 February 2025
Revised: 6 March 2025
Accepted: 10 March 2025
Published: 12 March 2025

Citation: Asafa, G.F.; Ren, S.; Mamun, S.S.; Gobena, K.A. DepthCloud2Point: Depth Maps and Initial Point for 3D Point Cloud Reconstruction from a Single Image. *Electronics* **2025**, *14*, 1119. https://doi.org/10.3390/electronics14061119

1. Introduction

Reconstructing 3D objects from a single input image is considered one of the most basic yet challenging problems in computer vision, while intuitively, humans can infer the structure of a scene and the shapes of objects within it from limited information. Even for regions that are highly occluded, we are able to guess a number of plausible shapes that could complete the object. This involves the intrinsic and distinctive human ability to reason on both missing and occluded parts from partial views and prior knowledge of geometric relations that allow seamless interaction with the environment. For example, imagine looking at a chair from a single angle; although parts of the chair, like its back legs, may be hidden, the human brain intuitively completes the missing structure based on past experience and understanding of shapes. The essence is that this has to be translated to machines and involves everything concerned with robotics, autonomous navigation, and augmented reality, where understanding the structure in 3D will play a major role in performing object manipulations and grasping and navigating unstructured environments [1,2].

Single-view 3D reconstruction remains a highly valuable yet challenging task due to the limited information available in a single image. Imagine scenarios where capturing multiple viewpoints of an object is impossible, for instance, analyzing a single historical photograph, reconstructing objects in dynamic scenes where only one image is available, or enabling robots to understand and interact with their environment based on a single camera feed. In such cases, multi-view methods, like structure-from-motion or triangulation, cannot be applied, leaving single-view 3D reconstruction as the only viable solution. This paper studies single-view 3D reconstruction to tackle these real-world challenges, aiming to infer 3D structures from minimal visual input while bridging the gap between what humans intuitively perceive and what machines can achieve [3,4]. Deep learning approaches have introduced data-driven solutions that allow models to infer 3D shapes from a single image by learning patterns and semantics from large datasets. Nonetheless, these techniques encounter difficulties when depth ambiguities occur. Consider observing an image of an airplane flying in the sky: in the absence of other background, it is challenging to ascertain whether the airplane is little and nearby or substantial and distant. This ambiguity in spatial perception poses difficulties in precisely reconstructing the three-dimensional structure. Likewise, occlusions, such as sections of the airplane obscured by clouds or other objects, generate voids in the input data, compelling the model to infer absent details. Due to these constraints and the fact that a single view only contains a limited amount of spatial information, reconstructions frequently lack precision or fine details [5,6].

The initial methodologies in deep learning for 3D reconstruction utilized voxel-based representations employing 3D convolutional neural networks (3D-CNNs) [7,8]. These approaches, however proficient in modeling volumetric data, encounter considerable obstacles. Voxel grids are resource-intensive, require substantial memory, and possess restricted resolution. Furthermore, voxel grids encompass internal characteristics of objects that are extraneous for the majority of reconstruction tasks, rendering them wasteful and frequently resulting in unsatisfactory results [9]. The aforementioned limitations have catalyzed the creation of more efficient representations, such as point clouds, which directly sample points on the visible surfaces of objects, thereby capturing critical geometric data without superfluous overhead. Point cloud-based methodologies, including PSGN [10] and Pixel2Point [11], have exhibited the advantages of point clouds in representing intricate geometry. These technologies provide a streamlined and effective substitute for voxel grids. Nevertheless, obstacles persist. The irregular distribution of points and the failure to maintain intricate object characteristics persistently impede the quality of reconstructions generated by these methods. To address these constraints, we utilize the benefits of depth maps and preliminary point clouds. We compute depth maps utilizing a pre-trained depth estimation network known as DPT [12] instead of depending on costly depth cameras. These estimated depth maps represent the relative distances of object surfaces, offering valuable geometric information essential for deducing 3D structures. In contrast to RGB photographs that solely convey appearance information, depth maps directly depict the third dimension, aiding in the resolution of ambiguity regarding object shapes. This is especially helpful in situations where spatial relationships are ambiguous due to occlusions or perspective distortions [13]. Initial point clouds enhance depth maps by acting as a structural prior, offering a foundational geometric depiction of an object. The starting point cloud, usually depicted as a uniform spherical distribution, directs the reconstruction process to generate uniformly dispersed points on the object's surface. This guarantees the preservation of finer details and helps the resulting point cloud closely match the actual shape of the object, even when input data are sparse or unclear [11,14].

In this work, we present DepthCloud2Point, a framework that combines the advantages of depth maps and preliminary point clouds to resolve depth ambiguities and compensate for absent spatial information, as shown in Table 1. Depth maps offer critical geometric data, whilst the structural prior derived from the initial point cloud guarantees consistent point distribution and intricate object reconstruction. Our approach effectively integrates these components, connecting sparse 2D inputs with precise, high-quality 3D reconstructions, therefore enhancing the field of single-view 3D reconstruction. The proposed architecture is meticulously crafted to tackle the difficulties of recreating 3D structures from a solitary 2D image. The process commences with the depth branch, which utilizes a pre-trained depth estimate network to generate depth maps. This phase is essential for retrieving spatial information frequently diminished during the projection of a 3D scene into 2D space. By assimilating these depth signals, the model acquires a fundamental comprehension of the object's shape. The encoder derives significant semantic and geometric characteristics from the input image. The encoder generates a comprehensive representation through consecutive convolutional layers, encapsulating both the visual and structural characteristics of the object. The retrieved features, enhanced with spatial insights from the depth branch, serve as the foundation for precise reconstruction. The concluding phase entails the generator, which produces a 3D point cloud by integrating the depth map, picture features, and a preliminary point cloud. The framework employs a concatenation technique to integrate geographical, semantic, and structural information into a cohesive representation. This guarantees a cohesive integration of features, enabling the generator to enhance the reconstruction with a consistent point distribution and maintained geometric details. This method enables the framework to convert sparse 2D inputs into precise and intricate 3D representations. This pipeline addresses the inherent challenges of single-view 3D reconstruction and outperforms current leading approaches, such as 3D-LMNet [15] and Pixel2Point [11]. The main contributions of this work can be summarized as follows:

1. We propose DepthCloud2Point, a framework that uses a CNN with depth maps, image features, and initial point clouds for single-view 3D reconstruction.
2. Rather than directly inferring the point cloud, we incorporate depth maps to offer comprehensive spatial cues for maintaining intricate object details and employ initial point clouds as geometric priors to fix the irregular point distribution, ensuring uniformity in point distribution.
3. The proposed model is evaluated on the ShapeNet and Pix3D datasets, demonstrating superior performance compared to state-of-the-art methods. Quantitative and qualitative results based on evaluation metrics such as Chamfer Distance (CD) and Earth Mover's Distance (EMD) highlight its effectiveness in achieving high-quality single-view 3D reconstructions.

The rest of this paper is organized as follows: In Section 2, we provide a review of related works, including single-view voxel-based 3D reconstruction and single-view point cloud-based 3D reconstruction. Then, in Section 3, we present the proposed methodology, including detailed information on the proposed model and loss function. To demonstrate the effectiveness of our proposed approach, extensive experiments and ablation analyses are conducted in Section 4. Finally, we conclude our work in Section 5.

Table 1. Limitations of voxel-based vs. point-cloud-based methods and how DepthCloud2Point addresses them.

Limitation	Voxel-Based Methods	Point Cloud-Based Methods	How DepthCloud2Point Solves It
Memory and Computation	Limited scalability due to volumetric grids.	Efficient, but suffers from irregular point distribution.	Uses depth maps and initial point clouds to optimize memory usage.
Representation Precision	Loss of details in voxelization.	Point clouds provide rich geometric details.	Enhances detail preservation through improved point distribution.
Occlusion and Depth Ambiguities	Struggles with resolving occlusions.	Depth ambiguity due to point distribution.	Addresses occlusions effectively using depth maps.
Fine Detail Preservation	Difficulty capturing fine details due to voxelization.	Allows for precise fine details to preserve distinct object details.	Captures fine details by leveraging depth information.

2. Related Work

2.1. Single-View Voxel-Based 3D Reconstruction

Voxel-based [16] representations were among the first approaches to 3D reconstruction. One of the fundamental works in this domain was by Choy et al. [9,17], who introduced a 3D convolutional neural network (3D-CNN) capable of generating volumetric outputs from a single image. Their encoder–decoder framework significantly improved 3D shape representation compared to earlier methods. However, their method suffered from low resolution due to voxel grid constraints, which limited the quality of the reconstructed shapes. Tulsiani et al. [18] further enhanced voxel reconstruction by proposing an unsupervised approach that leveraged multiple views with unknown camera poses. This method allowed for greater flexibility in reconstruction but was not well suited for single-view settings due to its reliance on multi-view inputs. Similarly, Shubham et al. [19] demonstrated that integrating additional 2D cues, such as depth maps and semantic segmentation, could enhance voxel reconstructions. However, voxel-based methods remained constrained by high memory consumption and inefficiency in capturing fine surface details due to their discretized grid structure, making them less practical for high-resolution reconstruction [5,9].

2.2. Single-View Point Cloud-Based 3D Reconstruction

To overcome the inefficiencies of voxel grids, point cloud-based models emerged as a more compact and surface-oriented alternative [9,20]. PSGN [10] was one of the earliest works in this domain, introducing a direct point cloud generation approach using neural networks. By bypassing volumetric representations, PSGN reduced memory requirements while improving reconstruction detail. However, its reliance on direct point generation led to inconsistencies in shape representation, as it lacked structural priors to guide the point cloud formation. RealPoint3D [21] addressed this limitation by combining 2D image features with 3D geometric priors. Using a dual-encoder architecture and priors extracted from ShapeNet, RealPoint3D improved the fidelity of generated point clouds. However, this approach relied on pre-existing 3D datasets, limiting its generalization to unseen objects. 3D-LMNet [15] introduced a two-stage framework that mapped 2D images into latent feature representations using a 3D autoencoder before synthesizing point clouds. While this method incorporated depth information, it struggled with generating uniform point distributions and suffered from incomplete reconstructions due to its multi-depth view dependency. Fan et al. [10] proposed a dense point cloud generation approach based on predicting predefined depth images and merging multiple depth maps to construct 3D

models. This method improved the uniformity of point distributions but required significant computational resources. More recent approaches, such as the work by Lu et al. [22], introduced a sparse-to-dense pipeline that first generated an initial sparse point cloud from RGB images and later refined it through densification stages, balancing efficiency and reconstruction quality. Pixel2Point [11] represented a major advancement by integrating CNN-extracted 2D image features with an initial spherical point cloud, which improved geometric fidelity. However, its sole reliance on image features limited its ability to generate uniformly distributed point clouds, making depth ambiguity a persistent issue. Recent implicit function-based methods, such as Occupancy Networks, have introduced continuous shape representations that achieve high-quality reconstructions. While these approaches excel in producing fine details, they often require complex optimization techniques and extensive computational resources, making them impractical for real-time applications.

DepthCloud2Point builds on the strengths of previous methods while addressing their limitations. It introduces a framework that combines depth maps, image features, and an initial point cloud to overcome the limitations of both voxel-based [5,9] and point cloud-based [22–25] methods. Unlike voxel-based approaches, which suffer from high computational costs and limited resolution, DepthCloud2Point generates point clouds directly from input images, ensuring lightweight yet detailed reconstructions. Compared to point cloud-based methods such as Pixel2Point and 3D-LMNet, it integrates depth maps as an additional source of spatial information, reducing the depth ambiguity inherent in single-view reconstructions. The initial spherical point cloud serves as a geometric prior, guiding the model toward more uniform and structured reconstructions. Unlike RealPoint3D, which relies on retrieving similar 3D models from a database, DepthCloud2Point generates point clouds without any dependency on external datasets. Additionally, it does not require auxiliary 2D supervision [26], such as silhouettes, to infer object structures. This simple yet robust design ensures more accurate, detailed, and efficient 3D reconstructions, addressing the key challenges in single-view 3D reconstruction.

3. Methodology

3.1. Architecture of DepthCloud2Point

The proposed architecture of DepthCloud2Point introduces a new pipeline that comprises depth maps, image features, and an initial point cloud to generate highly detailed and accurate 3D point clouds from a single 2D image, as illustrated in Figure 1. This provides a unique design given the limitation of traditional single-view reconstruction methods, which usually perform weakly in recovering depth information and generating uniformly distributed 3D points. It is an architecture that wraps a pre-trained depth estimator called DPT [12] at its very core and is able to predict a depth map from a given RGB image, thus providing the necessary spatial information that bridges the input from 2D to 3D. Image features are then extracted from these, merging into the backbone for reconstruction along with the geometrical structure of an initially estimated point cloud. Unlike traditional single-view approaches, which often rely on indirect heuristics for depth recovery, our method benefits from direct depth supervision via DPT, improving geometric consistency in the reconstructed 3D shape. To further optimize performance, the DPT model was fine-tuned on a dataset specific to our task. This fine-tuning process involved training on a curated subset of ShapeNet and Pix3D, ensuring the depth estimator aligns well with the geometry of the target objects. Since DPT-generated depth maps are used instead of depth cameras, the accuracy of depth estimation plays a crucial role in determining the quality of the final 3D reconstruction. To evaluate this, we compared our approach in Section 4. The results show that while our model effectively reconstructs objects using DPT-generated depth, reconstructions with depth maps exhibit higher structural accuracy. This is achieved

by segmenting the network into three large tasks: the depth map prediction, 2D feature extraction, and generating a point cloud. As shown in Algorithm 1, first the depth branch predicts a depth map from the given 2D image and hence provides the geometric context. Simultaneously, it extracts features with the encoder from both the depth map and the 2D image, thus capturing all visual and semantic information. The obtained features become integrated with the input point cloud as a geometric prior to an elaborate representation of the object. The merged data are then fed to the generator, which refines them and transforms them into the final 3D point cloud. Using the depth map, image features, and initial point cloud together, the architecture maximizes the potential for reconstructive performance in single-view 3D reconstruction and hence avoids some big challenges in reconstructive scenes: occlusions and incomplete spatial information. Each network part is described below.

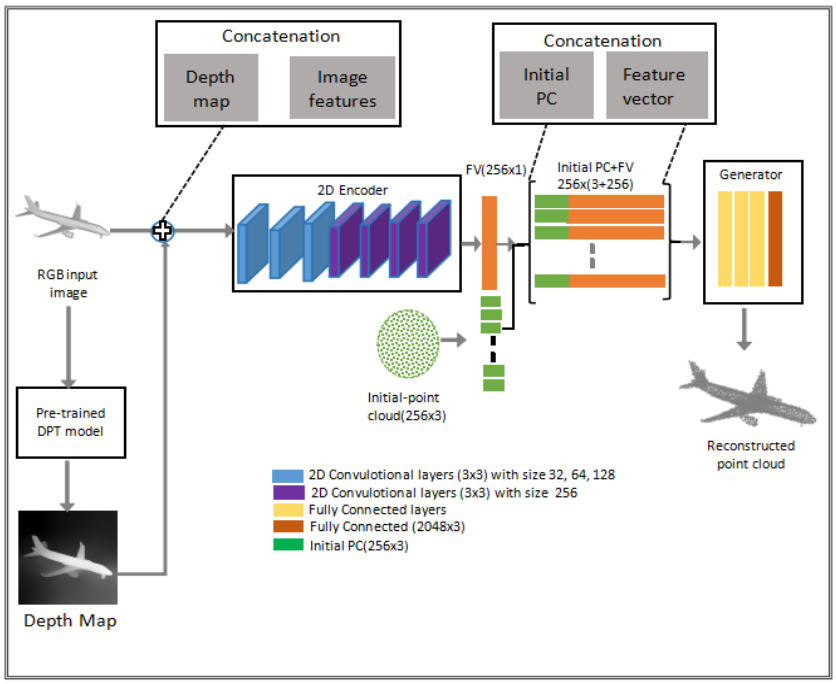

Figure 1. The proposed DepthCloud2Point architecture for 3D point cloud reconstruction.

3.1.1. Depth Map Branch

This depth branch constitutes a very significant aspect of the proposed architecture in 3D-CNN, as it provides the depth map that effectively carries important spatial information. Therefore, the designed branch basically utilized a pre-trained depth estimator called DPT that was tuned to take just one RGB image as an input and to return a depth map showing the relative distances of object surfaces. Estimating depth maps is one of the most challenging single-view 3D reconstruction issues since no explicit depth information is available in 2D images. In other words, by recovering this depth channel, it managed to elicit geometric context from the flat visual input through rearrangement, providing a backbone for reconstructing 3D structure. The first preprocessing step is resizing the input 2D image to a standard resolution of 384 × 384 pixels to bring it into coherence with the requirement of the depth estimator. This will ensure that most of the details relevant for the depth prediction are retained. This automatically results in an output depth map that will encode the object's spatial layout, including subtle curvatures, relative distances, and surface orientations. These are very important details that turn out to be useful in subsequential processes, hence enhancing the capability of the network to infer a realistic, coherent 3D structure. In return, this depth branch enriches the input data that the whole

reconstruction pipeline relies on, using geometrical clues, hence allowing a much more accurate and fine-grained 3D object reconstruction with the architecture suggested.

Algorithm 1 Algorithm of Proposed DepthCloud2Point Network

Require: Input image I, pre-trained depth model DPT, initial point cloud P_{initial}
Input: 2D RGB image
Output: Predicted 3D Point Cloud
1. Depth Map Prediction:
Resize input image I to 384×384 pixels
Compute depth map $D \leftarrow DPT(I)$
2. Feature Extraction Using Encoders:
Extract image features $F_{\text{image}} \leftarrow f_{\text{encoder}}(I)$
Extract depth features $F_{\text{depth}} \leftarrow f_{\text{encoder}}(D)$
3. Early Fusion (Merge Depth and Image Features):
Concatenate F_{image} and F_{depth} to form merged feature vector M
4. Initial Point Cloud Generation:
Generate initial point cloud P_{initial} with 256 uniformly distributed points
5. Merge Point Cloud and Features:
Concatenate P_{initial} with M to obtain final feature vector F_v
6. Point Cloud Generation Using Generator:
Generate $P_{\text{predicted}}$ of shape 2048×3
7. Output:
Return $P_{\text{predicted}}$ as the predicted 3D point cloud

3.1.2. Encoder

The encoder is a major component in the proposed 3D-CNN architecture designed to extract meaningful features from an input 2D image along with its depth map, laying the foundation for understanding an object's semantic and geometric properties. The encoder consists of seven 2D convolutional layers in succession, each followed by a ReLU activation function. This network architecture will progressively capture details, from low-level features like edges and textures to high-level abstractions of object parts and shapes. The first three convolutional layers produce feature maps of sizes 32, 64, and 128, while the remaining layers output 256 feature maps to refine the object representation. The kernel size for all layers is 3×3, and a stride of 2 is applied, enabling the model to downsample spatial dimensions while retaining critical information. Unlike pooling layers, these stridden convolutions are fully trainable in the context of this encoder; thus, it learns the best way of aggregating features. Once these convolution operations are performed, the feature extractors reshape them as compact information of size $1 \times 1 \times 256$ while preserving all the essential information in condensed form. In later stages of the pipeline, this reshaped feature vector is concatenated with the depth map and the initial point cloud, ensuring that the encoder provides both semantic and spatial information to the final 3D point cloud reconstruction. The extracted encoder features play a crucial role in reducing depth ambiguity by capturing both local and global contextual information, allowing the model to differentiate between objects at different depths more effectively. Additionally, by preserving fine details throughout the convolutional layers, the encoder ensures that small yet critical object structures, such as edges and contours, are retained in the final depth-aware representation. This refined feature extraction process not only strengthens the reconstruction's accuracy but also enhances the model's ability to infer spatially coherent and realistic 3D shapes.

3.1.3. Initial Point Cloud Providing Geometric Structure

The initial point cloud is an essential part of the 3D-CNN architecture proposed, as it offers a geometric prior with a structure for reconstruction. This spherical point cloud

contains 256 uniformly distributed points, which serve as the initial approximation of the 3D object. This prior helps the model generate a coherent 3D object with a uniform distribution, even from partial input data. To further enhance uniformity, the model leverages Chamfer Distance (CD) and Earth Mover's Distance (EMD) loss functions, which encourage smooth point distribution and prevent sparsity. Unlike voxel-based approaches that impose rigid grid structures, our method dynamically refines the point cloud by learning adaptive point placements. The model integrates the initial point cloud with features captured by the encoder and information represented in the depth map, ensuring that the geometric structure complements spatial and semantic information. This combination results in a robust representation that guides the reconstruction process, yielding accurate and uniformly distributed 3D point clouds while preserving intricate object details.

3.1.4. Generator

The generator of the 3D-CNN framework takes the fused (2D and 3D) feature representation and generates a high-resolution, articulated 3D point cloud. This is a lightweight model containing four fully connected layers that iteratively process until the output is produced. ReLU activation functions are used by the first three layers to assist the network in understanding complex, non-linear relationships and capture minute data about things. The generator outputs the reconstructed point cloud in a matrix format of 2048 × 3, with 2048 indicating the number of points and 3 denoting the dimensions x, y, and z. The output provides a comprehensive depiction of the object's overall form and its intricate characteristics, including curves and edges. The generator ensures that the 3D reconstruction is precise, consistent, and visually plausible by utilizing a combination of depth maps, encoder characteristics, and the initial point cloud. The model proposed is an end-to-end pipeline with its components. The depth branch first processes the 2D image input, which produces a depth map that helps to enrich the scene's spatial understanding. The encoder takes meaningful features from the image and depth map and merges the two with the original point cloud to form a feature vector. Finally, this vector is converted into the full 3D point cloud by the generator. This method solves problems with a single view, such as depth ambiguity and lack of spatial information. It is robust and efficient, making it one of the most interesting solutions in computer vision.

3.2. Loss Function

One of the very basic steps in training a CNN model involves specifying an appropriate loss function. This would depend on a variety of factors, including problem type, dataset characteristics, or the nature of the output values. Usually used to compute the difference between the model's predictions and the actual ground truth, the loss function should guide the weight correction of the model depending on the computed error. In this case, it specifically computes the ground truth's difference from the produced point cloud. For the loss function to effectively serve its purpose in the back-propagation process, there are some requirements to be met: it should, first of all, be computationally efficient and differentiable; secondly, it must be outlier-robust [10,27]. Since these requirements are met, the model will learn from this and further improve on the performance. The application of 3D supervision involves a loss function that quantitatively assesses the disparity between two 3D shapes: the predicted shape and the ground truth shape. An appropriate distance metric can be utilized to accomplish this task. In light of this, the necessary loss function L between two 3D shapes, $S^{pred}, S^{gt} \subseteq R^3$, is defined as:

$$L(s^{pre}, S^{gt}) = \sum d(s^{pred}, S^{gt}), \quad (1)$$

where S^{pred} and S^{gt} represent the predicted 3D shape and the ground truth, respectively. Given that the point cloud is an unordered representation, the loss function must be invariant to the arrangement of the points. We propose employing and comparing two different loss functions, Earth Mover's Distance (EMD) [28] and Chamfer Distance (CD) [29], to achieve this. The function could parallelize as a result, producing an output of superior quality. The higher the quality and correctness of the thing built, the lower the value found.

3.2.1. Chamfer Distance (CD)

CD is a method based on nearest-neighbor principles, which is characterized by its computational efficiency and adaptability for point sets of varying sizes. Let $X^{gt} \epsilon R^{Nx3}$ represent the ground truth and $X^{pred} \epsilon R^{Nx3}$ represent the generated point cloud, where N represents the number of points in the point cloud. The Chamfer Distance between X^{gt} and X^{pred} is defined as

$$d_{CD}(s_1, S_2) = \sum_{x \varepsilon S_1} \min_{y \varepsilon S_2} \|x - y\|_2^2 + \sum_{y \varepsilon S_1} \min_{x \varepsilon S_2} \|x - y\|_2^2, \quad (2)$$

where the point cloud represents the predicted and ground truth shapes sets S_1 and S_2 respectively, and x and y are two corresponding points that belong to S_1 and S_2, respectively. Nearest neighbor is the basis for the correspondence. Chamfer distance is piecewise smooth and continuous, with independent search processes at each position. This function produces excellent results and can be parallelized. Better and more precise forms are formed with lower values. The lack of a well-defined method to guarantee consistency in the generated point cloud is Chamfer distance's drawback. The reason for this is that the optimization process produces a minimum where a subset of points make up the complete form and the other points group together.

3.2.2. Earth Mover's Distance (EMD)

The Earth Mover's Distance (EMD) is a method employed to evaluate the disparity between two multi-dimensional distributions in a feature space, utilizing a distance metric for individual features referred to as the ground distance. EMD extends this distance from individual characteristics to include complete distributions.

$$d_{EMD} = \min_{\phi: s_1 \to s_2} \sum_{x \varepsilon s_1} \|x - \phi(X)\|_2, \quad (3)$$

where S_1, S_2 are the two point cloud sets of equal size ($|S_1| = |S_2|$) representing the predicted and ground truth shapes, and$\phi : S_1 \to S_1$ is a bijection. This one-to-one correspondence between points is based on an optimal assignment. Although EMD is differentiable and parallelizable, it is computationally expensive, particularly for high-resolution point clouds.

4. Experimental Analysis

4.1. Experimental Setup

Dataset: The approach proposed in this paper is evaluated on the ShapeNet [30] dataset and Pix3D [31] dataset. The ShapeNet dataset is a substantial, synthetic 3D dataset widely employed in 3D research, encompassing 3D model retrieval and reconstruction. It is an assemblage of 3D forms, depicted as textured CAD models categorized semantically. In our experiment, we specifically utilize the ShapeNetCore dataset, a refined and aligned subset of the ShapeNet dataset. This sample includes more than 50,000 distinct 3D models across 55 prevalent object categories. Every model has 24 photos captured at fixed angles. We concentrate on 13 categories and employ the 80%–20% train–test division as specified

by [30]. The Pix3D dataset comprises 7656 authentic photos along with associated metadata, including masks, ground truth CAD models, and position information. Pix3D is a publicly accessible dataset comprising matched pairings of real-world images and 3D models. It encompasses significant diversity in item shapes and backdrops, presenting considerable challenges. We evaluate and report the efficacy of the suggested technique on the chair, sofa, and table categories from the Pix3D dataset.

Implementation Details: The proposed DepthCloud2Point model was implemented in PyTorch [32], trained, and used because of the flexibility it offers for deep learning tasks. The input to the model is a resized 128 $\times$ 128 RGB image, which goes through the encoder to develop a compact latent feature representation of dimension 256. This is fused together with depth maps and an initial point cloud as the generator network outputs the final 3D point cloud of size 2048 $\times$ 3, where every point represents x, y, z coordinates in 3D space. As shown in Table 2, training was performed using the Adam optimizer [33], chosen because of its ability for adaptive learning rates, with a fixed learning rate of 5×10^{-5} and a fixed minibatch size of 32. The model was guided to generate complete and detailed 3D reconstructions in training through the use of CD and EMD as loss functions. It balances the computation efficiency and robust learning well enough to solve most of the single-view reconstruction challenges.

Table 2. Experimental parameters used in training and evaluation.

Parameter	Description
Batch Size	32
Learning Rate	0.00005
Optimizer	Adam
Loss Function	Chamfer Distance (CD), Earth Mover's Distance (EMD)
Number of Epochs	100
Hardware	NVIDIA RTX 3090
Dataset Used	ShapeNet, Pix3D
Dataset Split	80% Training, 20% Test
Evaluation Metrics	Chamfer Distance (CD), Earth Mover's Distance (EMD)

Baselines: The DepthCloud2Point model was evaluated on both synthetic and real-world data, and it was trained using the ShapeNet dataset. An ablation study has shown that depth maps and initial point clouds substantially enhance reconstruction accuracy, enabling the model to generate high-quality 3D point clouds from input images. The superior performance of the proposed model was corroborated through comparisons employing Chamfer Distance and Earth Mover's Distance against leading approaches, including 3D-LMNet [15] and Pixel2Point [11]. Moreover, experiments conducted on the Pix3D dataset demonstrated that the model surpassed existing methods regarding generalization across diverse real-world images.

4.2. Results and Discussion

This section presents a series of tests conducted to verify the effectiveness of the DepthCloud2Point for single-view 3D reconstruction on the ShapeNet dataset and Pix3D dataset. The results aim to answer the following research questions (RQs):

RQ1. *How well does the proposed model reconstruct 3D point clouds qualitatively from single-view input images across multiple object categories?*

The proposed model was tested using the ShapeNet test set, which comprises synthetically generated images showcasing one object from different views across 13 categories.

The qualitative results for eight categories, presented in Figure 2, include input images, depth maps, ground truth point clouds, and the model's predicted point clouds. The analysis shows that the generated point clouds match the ground truth really well, capturing important geometric and structural features of the objects effectively. The model was able to recreate complex details, like the slats on a chair's backrest and the gaps in between, showing that it can keep track of fine-level details. The model also did really well in creating slim and less typical parts, like the stretchers between chair legs, which are usually tough to recreate. This ability shows how well the model can adapt to complicated shapes.

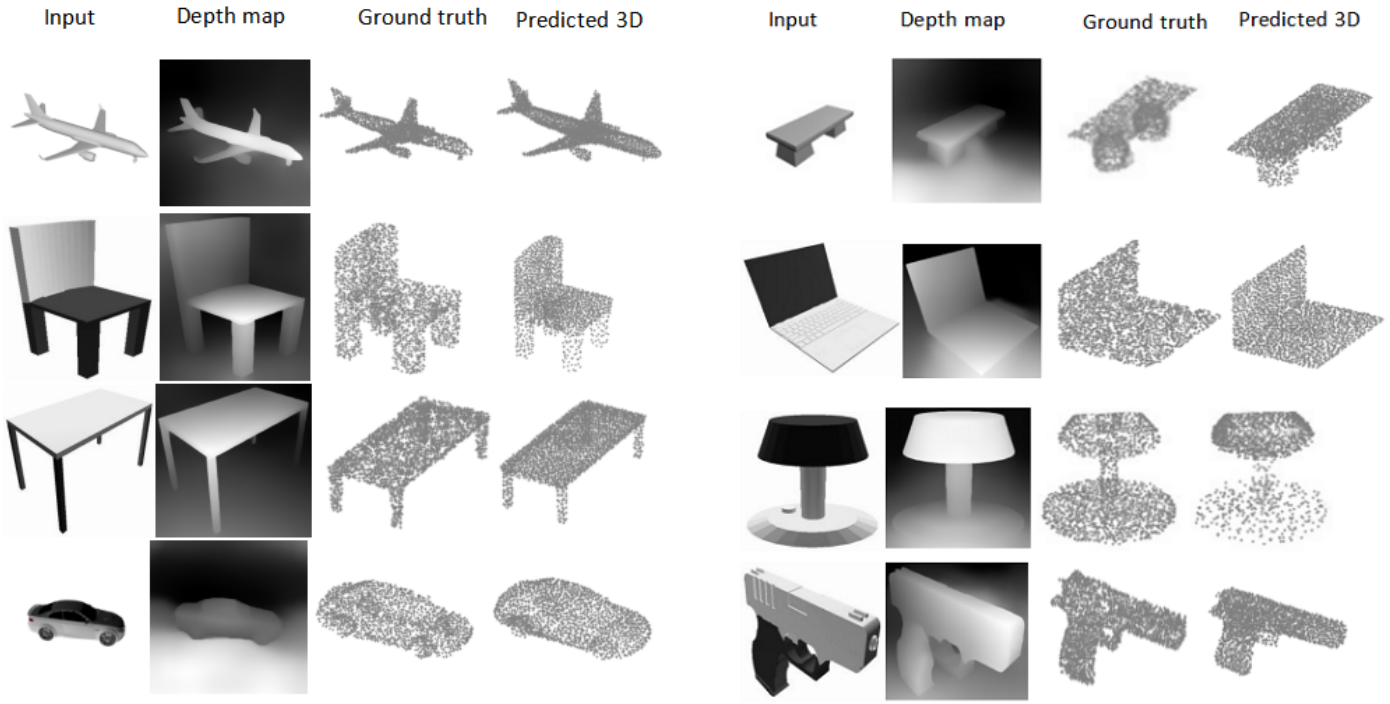

Figure 2. Qualitative results of ShapeNet on different categories.

The model consistently generated uniformly distributed point clouds that encompassed the entire surface of the objects, as illustrated in Figure 2. This uniformity guarantees accurate reconstructions and improves their similarity to the ground truth. Examples such as "laptop" and "gun" illustrate the model's capability to manage diverse geometrical configurations, accurately capturing intricate shapes and maintaining the finer details of complex structures. The capacity to generate spatially coherent and dense point clouds with realistic three-dimensional representations demonstrates the model's efficacy in executing 3D reconstruction from a single viewpoint. The qualitative performance demonstrates that the proposed model is a robust and reliable method for generating high-quality point clouds across various object categories.

RQ2. *How does the proposed model perform against state-of-the-art methods like 3D-LMNet and Pixel2Point on the ShapeNet dataset?*

The proposed model is evaluated in comparison to 3D-LMNet and Pixel2Point utilizing the ShapeNet dataset, maintaining identical training data and evaluation conditions to ensure fairness. The results presented in Table 3 indicate that the proposed model surpassed existing state-of-the-art methods across various metrics. The model demonstrated superior performance in 9 of 13 categories when evaluated with the Chamfer Distance, outperforming 3D-LMNet, and in 10 of 13 categories compared to Pixel2Point. In a similar evaluation using Earth Mover's Distance (EMD), the proposed model exceeded the perfor-

mance of 3D-LMNet across all 13 categories and outperformed Pixel2Point in 11 categories. The metrics highlight the model's accuracy in producing point clouds that closely align with the ground truth. In certain categories, alternative models demonstrated superior performance compared to DepthCloud2Point. The evaluation metrics further underscore the advantages of our approach. Earth Mover's Distance (EMD), measures point-to-point correspondence and penalizes mismatches. This metric highlights the high quality of our model's reconstructions, as they exhibit better alignment with the ground truth and successfully capture intricate object details.

Table 3. Quantitative comparison of single-view 3D point cloud reconstruction results on ShapeNet. All metrics are scaled by 100, and the bold metrics indicate better performance.

Category	CD			EMD		
	3D-LMNNet	**Pixel2Point**	**Ours**	**3D-LMNNet**	**Pixel2Point**	**Ours**
Airplane	3.34	3.29	**2.25**	4.77	3.82	**2.76**
Bench	4.55	4.59	**3.01**	4.99	4.31	**3.23**
Cabinet	6.09	**6.07**	6.19	6.35	4.94	**3.56**
Car	4.55	4.39	**3.45**	4.10	3.61	**2.45**
Chair	6.41	6.48	**3.90**	8.02	6.45	**5.03**
Lamp	7.10	6.58	**4.13**	15.80	8.45	**6.58**
Monitor	6.40	6.39	**4.21**	7.13	5.94	**4.42**
Rifle	**2.75**	2.89	2.82	6.08	**4.25**	4.53
Sofa	5.85	5.85	**4.78**	5.65	5.03	**3.82**
Speaker	**8.10**	8.39	8.45	9.15	**7.37**	7.43
Table	6.05	6.26	**4.89**	7.82	6.05	**5.15**
Telephone	4.63	**4.27**	4.33	5.43	3.77	**2.65**
Vessel	**4.37**	4.55	4.45	5.68	4.89	**4.06**
Mean	5.40	5.38	**4.37**	7.00	5.30	**4.28**

Similarly, the Chamfer Distance, which measures the nearest-point correspondence in both directions (generated-to-ground truth and ground truth-to-generated), affirms the precision of our model. Unlike prior methods, it achieves dense and evenly distributed point clouds without requiring point count equalization. In some categories, other models outperformed DepthCloud2Point. For instance, Pixel2Point achieved the lowest Earth Mover's Distance (EMD) for the rifle category (4.25), demonstrating its superior capability to align and distribute points for highly detailed, elongated structures. In the vessel category, 3D-LMNet performed better in terms of Chamfer Distance (4.37), suggesting its robustness in reconstructing objects with smooth, curved surfaces. Additionally, for the cabinet category, Pixel2Point also exhibited better results in terms of Chamfer Distance (6.07), reflecting its strength in reconstructing large, box-like objects. These results highlight specific strengths of the competing models in handling particular geometries or surface characteristics. While our method demonstrates overall superior performance, these cases indicate areas for potential improvement in reconstructing certain object types. The advantages of the proposed model stem from its integration of depth maps and initial point clouds as inputs, which provide a richer geometric prior compared to the two-stage approach of 3D-LMNet or the CNN-based method of Pixel2Point. Unlike 3D-LMNet, which often produces sparse and unordered point clouds due to inconsistencies in point fusion, the proposed model generates dense and evenly distributed reconstructions. Similarly, it avoids the clustering and incomplete reconstructions observed with Pixel2Point. The quantitative results affirm that the proposed model achieves higher performance across various object categories, showcasing its robustness, precision, and ability to overcome the limitations of prior methods.

As shown in Figure 3, there are qualitative comparisons indicating the advantages of our method. Our model reconstructs fine-level details such as the gripping and the magazine when reconstructing the rifle. The 3D-LMNet output totally lacks these features, while Pixel2Point presents them vaguely through fused and poorly separated parts. Our model produces point clouds that do not cluster like Pixel2Point or have missing or far points like 3D-LMNet. The point clouds produced are uniformly and evenly spread over the entire shape of the object. Through the fused use of depth maps and an initial point cloud, the model deals with the issues of existing methods. It always makes reconstructions that are more accurate and detailed but also overall more uniformly shaped and coherent. Our method proves effective as it can faithfully reproduce small features while maintaining geometric fidelity. In the end, our model proves to be a large step forward, which offers a robust and efficient single-view 3D reconstruction solution.

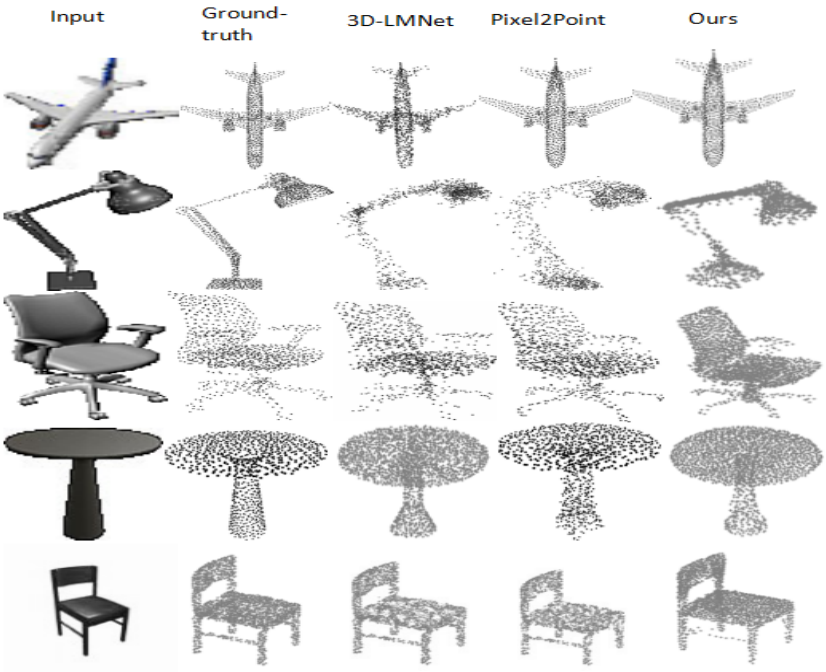

Figure 3. Comparison results between different methods on ShapeNet.

RQ3. *How well does the proposed model generalize to real-world data, such as the Pix3D dataset?*

To evaluate the model's generalization capabilities, experiments were conducted on the Pix3D dataset [31], which includes real-world images with object masks, accurate CAD models, and pose information. The dataset was preprocessed by cropping the images and masking background regions to ensure uniform evaluation. Results presented in Table 4 indicate that the proposed model achieved lower Chamfer Distances and EMD in two out of three categories (chair and table) compared to 3D-LMNet and Pixel2Point, while performing slightly less effectively in the sofa category. For the sofa category, Pixel2Point achieved the best Chamfer Distance (CD) of 3.95 and the best Earth Mover's Distance (EMD) of 3.28, highlighting its ability to handle real-world objects with curved and irregular surfaces better than our model. The gap may arise from the distinctive geometric and structural attributes of sofas, which often include wide, flat surfaces with fewer pronounced edges or details. These features may not completely utilize the advantages of depth maps and structural priors, thus constraining the model's capacity to attain optimal reconstructions in this domain. Despite this, the model demonstrates its ability to handle real-world data effectively, maintaining high-quality reconstruction performance across other categories. Qualitative results shown in Figure 4 further validate the model's generalization capabili-

ties. The reconstructions produced by the proposed model closely resemble the geometry of real objects, surpassing the outputs of 3D-LMNet and Pixel2Point in terms of coherence and structural accuracy. While 3D-LMNet often generated sparse and incomplete reconstructions, and Pixel2Point exhibited fused or poorly separated parts, the proposed model successfully preserves high-level details and structural features. These results highlight the model's ability to bridge the gap between synthetic and real-world data, making it a reliable choice for practical applications in 3D reconstruction. All in all, the results of the Pix3D evaluation prove the effectiveness of the proposed approach compared to other methods, especially when it comes to real-life data. By using geometric and structural priors to depth maps and initial point clouds, our method not only closes the gap between synthetic and real but also offers better quality in 3D reconstruction.

Table 4. Single-view 3D reconstruction in the real-world Pix3D dataset. All metrics are scaled by 100, and the bold metrics indicate better performance.

Category	CD			EMD		
	3D-LMNNet	**Pixel2Point**	**Ours**	**3D-LMNNet**	**Pixel2Point**	**Ours**
Chair	7.35	6.82	**3.79**	9.14	7.45	**5.23**
Sofa	8.18	**3.95**	6.75	7.22	**3.28**	6.12
Table	11.20	5.22	**3.70**	12.73	5.17	**4.45**
Mean	8.91	5.33	**4.75**	9.70	5.30	**5.16**

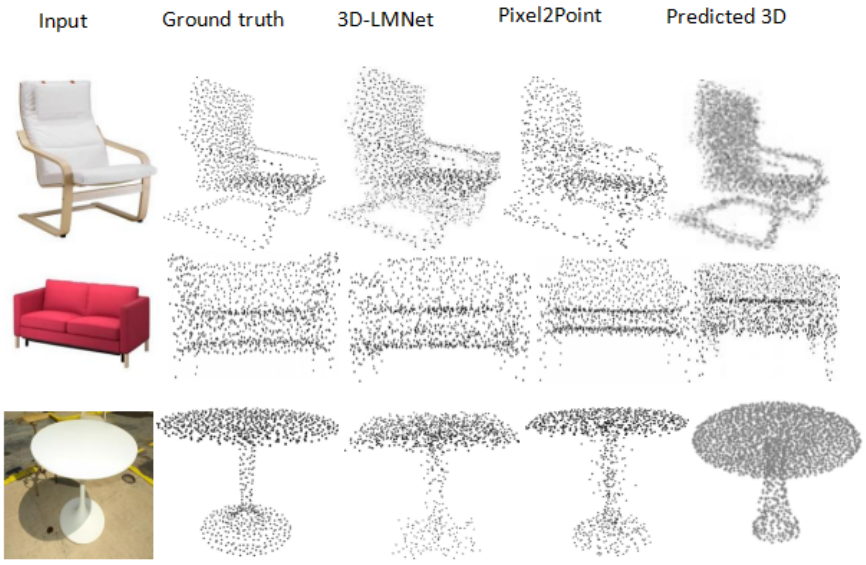

Figure 4. Qualitative results on chair, sofa, and table categories from Pix3D dataset.

Even while DepthCloud2Point shows good generalization on real-world data from Pix3D, the quality of reconstruction may still be impacted by specific issues specific to real-world images. In contrast to synthetic datasets, Pix3D includes complex backdrops, partial occlusions, and a range of lighting situations, all of which add noise and missing data to depth estimation. Our approach captures fine geometric details like synthetic data, maintaining good reconstruction fidelity when the object is well-lit and separated from the backdrop. Nevertheless, the depth map estimation may become less accurate in situations with strong illumination contrasts or shadows, which could result in distortions in the rebuilt point cloud. Additionally, background clutter and occlusions introduce ambiguities in spatial information. When objects are partially occluded, such as a chair behind a table, DepthCloud2Point attempts to infer the missing geometry, but reconstructions may exhibit

incomplete or imprecise structures. Despite these challenges, the model remains robust in maintaining global shape consistency, as it leverages geometric priors from the initial point cloud to regularize the reconstruction. These findings suggest that DepthCloud2Point is well suited for controlled environments where objects are clearly visible but may require additional refinement techniques for deployment in highly cluttered or occluded scenes.

4.3. Ablation Studies

Effect of the Depth Map and Initial Point Cloud

In order to quantify the contribution of depth maps and initial point clouds toward point cloud reconstruction, two ablated versions of the proposed model were considered. The first configuration, depicted in Figure 1, uses both a depth map and an initial point cloud, while the second configuration omits these elements and is solely based on the input image to produce the point cloud. Both architectures were trained on ShapeNet training and tested on ShapeNet validation. The results, shown in Figure 5, demonstrate a clear difference between the two settings in reconstruction quality. The model failed to generate correct and uniformly distributed point clouds without depth maps and initial point clouds. This resulted in inappropriate point distributions on object surfaces and poor reconstruction of delicate details. For example, chairs reconstructed without these components exhibited clumps of points at the back corners and poorly defined legs. Similarly, the tables did not have clearly defined legs, and the aircraft reconstructions lacked any critical features such as engines and tails, with points overly concentrated along the fuselage. This further brings out the inability of the model to capture structural details and distribute points evenly in the absence of geometric priors.

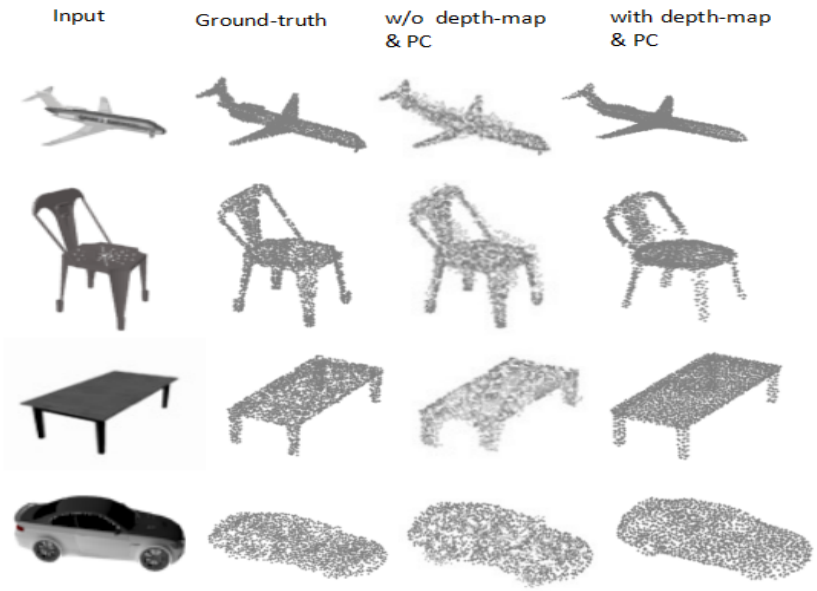

Figure 5. Qualitative results of the different setups of the proposed model on ShapeNet.

By contrast, depth maps and initial point clouds greatly improved reconstruction quality. With this input, models generated well-distributed point clouds on object surfaces, accurately reconstructing thin, articulated structures. For instance, the model showed thin chair and table legs and detailed features such as airplane engines and tails, with high reconstruction quality. These results point out the advantage of using a combination of depth maps and initial point clouds, which give both geometric information and structural priors to guide the model toward accurate and detailed reconstructions. Quantitative results are summarized in Table 5 and confirm the effectiveness of the proposed components.

The introduction of depth maps and initial point clouds significantly improved both Chamfer and EMD, indicating higher precision and completeness of the reconstructed point clouds. Depth maps provide critical geometric information, while initial point clouds provide structural priors. These ensure that the model generates dense and uniformly distributed points and recovers the fine details of objects. Without them, the model's reconstructions are less dense, less uniform, and less detailed, underlining their crucial role in guiding the model toward high-quality 3D reconstruction. In summary, depth maps and initial point clouds will play a vital role in ensuring that the 3D reconstructions are accurate, fine in detail, and well distributed. The model is, therefore, able to overcome the weakness of purely image-based methods through its use, thereby yielding reconstructions truly representative of the geometrical and fine details of objects.

Table 5. Quantitative comparison of different setups of the proposed model on ShapeNet. All metrics are scaled by 100, and the bold metrics indicate better performance.

Category	CD		EMD	
	w/o **Depth & PC**	**With Depth & PC**	*w/o* **Depth & PC**	**With Depth & PC**
Airplane	4.65	**2.25**	5.92	**2.76**
Bench	5.74	**3.01**	7.23	**3.23**
Cabinet	**5.82**	6.19	4.12	**3.56**
Car	6.32	**3.45**	6.13	**2.45**
Chair	7.11	**3.90**	11.16	**5.03**
Lamp	6.07	**4.13**	8.13	**6.58**
Monitor	7.22	**4.21**	9.32	**4.42**
Rifle	3.45	**2.82**	10.46	**4.53**
Sofa	6.32	**4.78**	4.14	**3.82**
Speaker	**7.12**	8.45	8.12	**7.43**
Table	6.21	**4.89**	7.23	**5.15**
Telephone	**4.14**	4.33	3.12	**2.65**
Vessel	5.00	**4.45**	7.51	**4.06**
Mean	5.76	**4.37**	7.12	**4.28**

4.4. Failure Cases

The proposed model fails to generate very accurate shapes in some cases. Figure 6 shows some failure cases. Most of the thinner and narrower parts of the objects are missed, such as the airplane's tail fins and the finer structural details of the modern chair. The model also struggles with objects that have additional structural complexity, such as the van's wheels and side mirrors, which are either deformed or missing. Similarly, the airplane's thin components, including the wingtips, are not well captured, leading to a less defined shape. Normally, these parts are omitted because the network does not fully capture their significance during reconstruction. However, in cases where the model encounters such challenges, it attempts to generate the closest possible shape to the input. For instance, in the case of the basket, the model reconstructs a plausible structure, but the finer grid-like details are lost, resulting in a more simplified and smoothed shape.

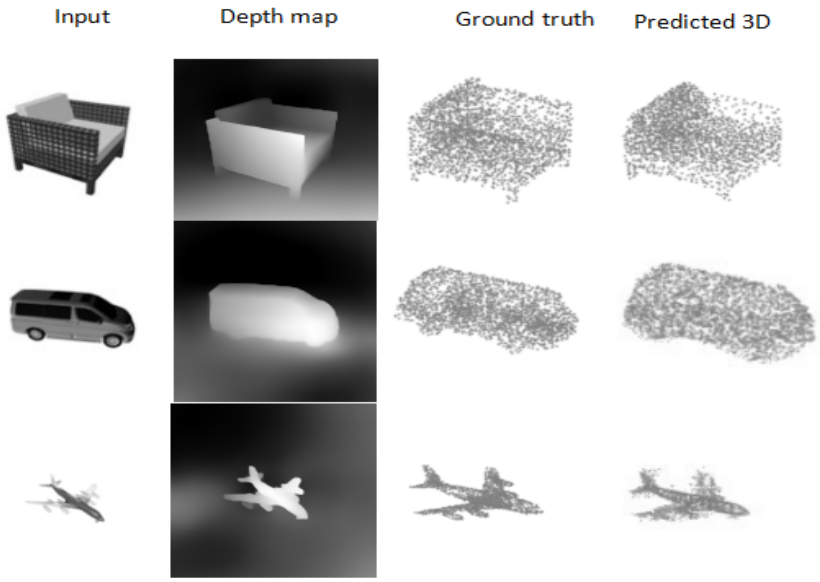

Figure 6. Failure cases of our method on the ShapeNet dataset.

5. Conclusions

In this paper, we propose DepthCloud2Point, a novel framework for accurate 3D reconstruction from a single RGB image. Unlike previous classical single-image reconstruction methods, which always suffer from spatial ambiguities, DepthCloud2Point combines depth maps and RGB images to generate an initial point cloud, which is further refined by the generator network to produce high-quality 3D reconstructions. Our method outperforms state-of-the-art methods on both the ShapeNet and Pix3D datasets, proving efficient in-depth and visual feature combinations to enhance spatial understanding and reconstruction quality. The model has promising applications in robotics for object recognition and manipulation, medical imaging for anatomical reconstruction, industrial quality control for defect detection, and cultural heritage preservation for non-invasive artifact restoration. Despite its advancements, DepthCloud2Point faces challenges in handling occluded or highly complex structures where feature extraction and point cloud refinement may lose precision. Future research should focus on optimizing feature integration strategies, improving model efficiency, and adopting more advanced architecture to improve reconstruction quality. We believe that DepthCloud2Point represents a significant step toward advancing single-view 3D reconstruction and provides a strong foundation for further research and practical applications.

Author Contributions: Conceptualization, G.F.A. and S.R.; methodology, G.F.A. and S.R.; software, G.F.A.; validation, G.F.A., S.S.M. and S.R.; formal analysis, G.F.A. and S.R.; investigation, S.R.; resources, G.F.A. and S.R.; data curation, G.F.A. and K.A.G.; writing—original draft preparation, G.F.A.; writing—review and editing, S.R., S.S.M. and K.A.G.; visualization, K.A.G. and S.S.M.; supervision, S.R.; project administration, G.F.A. 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 on 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. Ehsani, K.; Han, W.; Herrasti, A.; VanderBilt, E.; Weihs, L.; Kolve, E.; Kembhavi, A.; Mottaghi, R. Manipulathor: A framework for visual object manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 4497–4506.
2. Kent, D.; Behrooz, M.; Chernova, S. Construction of a 3D object recognition and manipulation database from grasp demonstrations. *Auton. Robot* **2016**, *40*, 175–192. [CrossRef]
3. Placed, J.A.; Strader, J.; Carrillo, H.; Atanasov, N.; Indelman, V.; Carlone, L.; Castellanos, J.A. A survey on active simultaneous localization and mapping: State of the art and new frontiers. *IEEE Trans. Robot.* **2023**, *39*, 1686–1705. [CrossRef]
4. Nousias, S.; Lourakis, M.; Bergeles, C. Large-scale, metric structure from motion for unordered light fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3292–3301.
5. 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/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3405–3414.
6. Henderson, P.; Ferrari, V. Learning single-image 3D reconstruction by generative modelling of shape, pose and shading. *Int. J. Comput. Vis.* **2020**, *128*, 835–854. [CrossRef]
7. Deng, J.; Shi, S.; Li, P.; Zhou, W.; Zhang, Y.; Li, H. Voxel R-CNN: Towards high performance voxel-based 3D object detection. *AAAI Conf. Artif. Intell.* **2021**, *35*, 1201–1209. [CrossRef]
8. Tang, H.; Liu, Z.; Zhao, S.; Lin, Y.; Lin, J.; Wang, H.; Han, S. Searching efficient 3D architectures with sparse point-voxel convolution. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 685–702.
9. Fahim, G.; Amin, K.; Zarif, S. Single-View 3D reconstruction: A survey of deep learning methods. *Comput. Graph.* **2021**, *94*, 164–190. [CrossRef]
10. 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.
11. 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]
12. Ranftl, R.; Bochkovskiy, A.; Koltun, V. Vision transformers for dense prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 19–25 June 2021; pp. 12179–12188.
13. Li, Z.; Wang, X.; Liu, X.; Jiang, J. Binsformer: Revisiting adaptive bins for monocular depth estimation. *IEEE Trans. Image Process.* **2024**, *33*, 3964–3976. [CrossRef] [PubMed]
14. Sohail, S.S.; Himeur, Y.; Kheddar, H.; Amira, A.; Fadli, F.; Atalla, S.; Copiaco, A.; Mansoor, W. Advancing 3D point cloud understanding through deep transfer learning: A comprehensive survey. *Inf. Fusion* **2024**, *9*, 110–121. [CrossRef]
15. Mandikal, P.; Navaneet, K.L.; 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.
16. Liu, F.; Liu, X. Voxel-based 3D detection and reconstruction of multiple objects from a single image. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 2413–2426.
17. Choy, C.B.; Xu, D.; Gwak, J.Y.; Chen, K.; Savarese, S. 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 628–644.
18. Tulsiani, S.; Efros, A.A.; Malik, J. Multi-view consistency as supervisory signal for learning shape and pose prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2897–2905.
19. Tulsiani, S.; Zhou, T.; Efros, A.A.; Malik, J. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2626–2634.
20. Lu, Y.; Wang, S.; Fan, S.; Lu, J.; Li, P.; Tang, P. Image-based 3D reconstruction for Multi-Scale civil and infrastructure projects: A review from 2012 to 2022 with new perspective from deep learning methods. *Adv. Eng. Inform.* **2024**, *59*, 102268. [CrossRef]
21. Xia, Y.; Wang, C.; Xu, Y.; Zang, Y.; Liu, W.; Li, J.; Stilla, U. RealPoint3D: Generating 3D point clouds from a single image of complex scenarios. *Remote Sens.* **2019**, *11*, 2644. [CrossRef]
22. Lu, F.; Chen, G.; Liu, Y.; Zhan, Y.; Li, Z.; Tao, D.; Jiang, C. Sparse-to-dense matching network for large-scale LiDAR point cloud registration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 11270–11282. [CrossRef] [PubMed]
23. Mescheder, L.; Oechsle, M.; Niemeyer, M.; Nowozin, S.; Geiger, A. Occupancy networks: Learning 3D reconstruction in function space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4460–4470.
24. Guo, Y.; Wang, H.; Hu, Q.; Liu, H.; Liu, L.; Bennamoun, M. Deep learning for 3D point clouds: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 4338–4364. [CrossRef] [PubMed]

25. Wu, W.; Qi, Z.; Li, F. PointConv: Deep convolutional networks on 3D point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9621–9630.
26. Pan, M.; Liu, J.; Zhang, R.; Huang, P.; Li, X.; Xie, H.; Wang, B.; Liu, L.; Zhang, S. RenderOcc: Vision-centric 3D occupancy prediction with 2D rendering supervision. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 13–17 May 2024; pp. 12404–12411.
27. Shen, W.; Jia, Y.; Wu, Y. 3D shape reconstruction from images in the frequency domain. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4471–4479.
28. Zhang, Z.; Zhang, Y.; Zhao, X.; Gao, Y. EMD metric learning. *AAAI Conf. Artif. Intell.* **2018**, *32*, 4490–4497. [CrossRef]
29. Wu, T.; Pan, L.; Zhang, J.; Wang, T.; Liu, Z.; Lin, D. Balanced Chamfer distance as a comprehensive metric for point cloud completion. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 29088–29100.
30. 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.
31. 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, 18–22 June 2018; pp. 2974–2983.
32. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8024–8035.
33. Zhang, Z. Improved Adam optimizer for deep neural networks. In Proceedings of the 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS), Banff, AB, Canada, 4–6 June 2018; pp. 1–2.