


Structure learning for 3D Point Cloud Generation from Single RGB Images

T. Ben Charrada ¹ , H. Laga ²  and H. Tabia ³

¹ Reezocar, France. ² Murdoch University, Australia. ³ IBISC, Univ. Evry, Université Paris-Saclay, France

Abstract

3D point clouds can represent complex 3D objects of arbitrary topologies and with fine-grained details. They are, however, hard to regress from images using convolutional neural networks, making tasks such as 3D reconstruction from monocular RGB images challenging. In fact, unlike images and volumetric grids, point clouds are unstructured and thus lack proper parameterization, which makes them difficult to process using convolutional operations. Existing point-based 3D reconstruction methods that tried to address this problem rely on complex end-to-end architectures with high computational costs. Instead, we propose in this paper a novel mechanism that decouples the 3D reconstruction problem from the structure (or parameterization) learning task, making the 3D reconstruction of objects of arbitrary topologies tractable and thus easier to learn. We achieve this using a novel Teacher-Student network where the Teacher learns to structure the point clouds. The Student then harnesses the knowledge learned by the Teacher to efficiently regress accurate 3D point clouds. We train the Teacher network using 3D ground-truth supervision and the Student network using the Teacher's annotations. Finally, we employ a novel refinement network to overcome the upper-bound performance that is set by the Teacher network. Our extensive experiments on ShapeNet and Pix3D benchmarks, and on in-the-wild images demonstrate that the proposed approach outperforms previous methods in terms of reconstruction accuracy and visual quality.

Keywords: CNN, Deep learning, 3D parameterization

CCS Concepts

• Computing methodologies → Point-based models;

1. Introduction

of the 3D point clouds. 3D object reconstruction from monocular RGB images is an important task, which is fundamental for augmented reality, mixed reality, holoportation, and metaverse [WCL19, LCRCN22, PTHPS20, ASY22, Tan22, AB22, SKKP22]. Despite being extensively investigated by the computer graphics, vision, and machine learning communities, real-time accurate 3D reconstruction of complex 3D objects of arbitrary topologies remains a challenge due to the complex structure of real-world objects and the high computation requirements of the state-of-the-art methods.

State-of-the-art image-based 3D reconstruction pipelines [HLB21] can be classified based on how 3D shapes are represented. Voxel-based methods [CXG*16, WWX*17, TZEM17, SWZ*18] can represent, and thus reconstruct, objects of arbitrary topologies. Due to their reliance on volumetric grids and 3D convolutions, these methods are expensive in terms of computation time and memory requirements. While neural implicit representations address the memory issue, they remain expensive at runtime.

Mesh-based methods [WZL*18, KUH18, KTEM18, PHC*19] are lighter but rely on template deformation. Thus, they can only reconstruct objects that have the same topology as the template. Ben Charrada *et al.* [BCTCL22] have proposed a method to address the topology issue in mesh-based methods by simultaneously deforming the geometry and topology of an initial template using face pruning operations. However, these operations can result in non-water tight meshes, which can significantly decrease the accuracy and quality of the reconstruction. Point-based representations [FSG17, LKL18, JSQJ18, MMAB18, GWM18, LZZ*19, SWL*18, LPZR18, LGCR19, ZKG18, MR19, WSL18, PEW21] are compact and flexible. They can be used to represent complex geometric and topological structures and fine geometric details. They, however, lack the notion of structure, or parameterization, and thus, are hard to regress from images using modern deep networks, which rely on convolutional operations that require grid-structured data. It is worth noting that, in this paper, we use the terms *parameterization* and *structure* interchangeably. They both refer to the process of structuring point clouds through parametrization.

In this paper, we propose a novel network architecture and a novel training strategy to reconstruct accurate 3D point clouds from monocular RGB images. Our key observation is that decomposing the 3D reconstruction problem into two sub-problems, one for structure, or parameterization, recovery and another for the geometry reconstruction once the structure is known, makes the learning task drastically easier. Thus, we propose a novel architecture that is composed of two sub-networks: **(1) a Teacher network** that learns how to structure 3D point clouds. It receives as input the ground truth 3D point cloud and a set of grid-structured 2D points. It then uses folding operations [YFST18] to deform the 2D points into the 3D space to fit the input 3D point cloud. It outputs the 3D reconstruction of the input ground-truth point cloud as well as the latent representation that encodes the points. This initial 3D reconstruction, which is parameterized by the input 2D grid, will then supervise, through a novel structural loss term, the training of the second, referred to as Student, network to ensure that the correct structure is reconstructed when regressing the 3D geometry. **(2) A Student network** whose role is to recover the 3D geometry by mapping RGB images to the data generated by the Teacher network. Since the latter is structured, the Student network's task becomes much easier than traditional point-based methods that learn how to regress unstructured point clouds from input images. In this paper, the terms *Teacher* and *Student* refer to our approach where one network supervises another. Note that this terminology is distinct from the context of knowledge distillation.

Our second key observation is that although end-to-end training is appealing, it is challenging in practice given the large number of parameters the network needs to learn. We show that better reconstruction can be obtained by training separately the Teacher and the Student networks. While we train the Teacher network using the Chamfer Distance (CD), we train the Student network using a novel *Latent Distance* loss, which, instead of directly comparing point clouds as done with the CD, compares their projection onto the latent space of the Teacher network. This addresses an important limitation of the CD, which is its saturation and insensitivity to subtle changes that can lead to non-optimal convergence. Finally, we employ an adversarial network to further refine the reconstructed 3D objects and overcome the upper-performance limit set by the Teacher network. We train the Teacher network end-to-end using the Chamfer Distance and a novel downsampling regularization method to remove instance-specific details, preventing the network from overfitting.

In summary, this paper makes the following contributions: **(1)** A novel Teacher-Student architecture and a novel training methodology that decouple the 3D geometry reconstruction problem from the structure understanding problem. **(2)** A novel loss function termed *Latent Distance* as an alternative to the widely used Chamfer loss. **(3)** A novel refinement stage that relies on an adversarial loss to address the limitation of the Chamfer distance. **(4)** We also show that the proposed framework can also be applied to 3D mesh reconstruction, and show that it outperforms state-of-the-art methods in terms of reconstruction accuracy. The remaining of this paper is organized as follows; Section 2 reviews the related work. Section 3 describes in detail the proposed method. Section 4 describes the loss functions used to train the proposed framework.

Section 5 presents the results and compares the performance of our proposed method to the state-of-the-art. We conclude in Section 6.

2. Related work

Deep learning methods, which revolutionized the field of image-based 3D reconstruction, train neural networks to infer the 3D geometry and structure of objects directly from a monocular RGB image. While the focus of our study is on single-view paradigms, we acknowledge that complementary techniques have emerged alongside. Multi-view reconstruction techniques such as Pixel2Mesh++ [WZC*22] capitalize on multiple image perspectives to construct detailed 3D models. Diffusion-based models, such as MeshDiffusion [LFB*23], shift the paradigm by focusing on the generation of realistic 3D shapes using meshes. Another avenue to consider is primitive-based reconstruction techniques such as IMS2Truct, which break down complex geometries into basic primitives, simplifying the overall reconstruction process [NLX18]. In the following sections, we will classify the state-of-the-art based on the 3D representation they use. For an introduction and comprehensive overview, readers are referred to the surveys by Han et al. and Laga et al. [HLB21, LJB20].

Volumetric methods use 3D voxel grids such as occupancy maps [LGOA18, TEM18] and (truncated) Signed Distance Functions (SDF) [DRQN17, CSO*18, KLR18, CLK*18], and extend the traditional 2D convolutions used on images to 3D. While convenient, they are very expensive in terms of memory requirements. Thus, they are limited to low-resolution reconstructions and are not suitable for edge devices. Several methods, e.g., [WLG*17, TDB17, RUG17, LXC*17, HTM19], proposed to use space partitioning techniques to reduce the memory requirements of volumetric methods. However, these methods require learning both the octree structure as well as its content (i.e., the 3D geometry) and thus result in complex network architectures.

Methods that use neural implicit functions [PFS*19, MON*19, LWL20, LZ21, DP22, AB22] address the resolution issue faced by volumetric methods. They are efficient to train and allow the reconstruction of 3D objects at an arbitrary resolution. They are, however, expensive at runtime since, to extract the mesh, one needs to evaluate the implicit function at every 3D location of a discretization 3D grid, which is then fed to a Marching Cube algorithm to extract the mesh.

Mesh-based representations that rely on deforming a canonical template to fit the target 3D geometry [WZL*18, MBM*17, PHC*19, HWX*21] cannot reconstruct objects that have a topological structure that is different from the template. Some of the recent works attempted to address this problem through topology modification by face pruning [PHC*19, BCTCL22]. For example, TopoNet [BCTCL22] addressed the topology issue by simultaneously deforming the geometry and the topology of an initial template. It introduced a face-pruning stage that is based on a reinforcement learning network. Pruning operations, however, can result in meshes that are not watertight. Other methods use an intermediate volumetric representation to recover the topology, followed by mesh refinement using Graph CNNs [GMJ19, THP*19]. Due to the computational cost, the intermediate volumetric representation is

usually of low resolution. For instance, Tang *et al.* [THP*19] use grids of size 128^3 while Gkioxari *et al.* [GMJ19] attempted to reduce the computational cost by using grids of size 48^3 . These methods also suffer from inconsistencies in the orientation of the surface normals, and require a refinement step, which brings an additional complexity. Thus, a trade-off between reconstruction accuracy and visual aspect is often needed.

Point-based representations are light and flexible but are unstructured. The success of PointNet and its variants [QSMG17, QYSG17] made processing point clouds without relying on intermediate representations possible. Subsequently, several methods proposed to apply convolutional operation on point clouds [TQD*19, BPM20, XFX*18, LBS*18, WQF19]. Also, Fan *et al.* [FSG17] introduced the Chamfer Distance (CD) and the Earth Mover Distance (EMD) as loss terms between the ground truth and the reconstructed point clouds.

Our hypothesis in this paper is that the problem of point-based reconstruction can be rendered easier by structuring the point cloud, allowing lighter networks to generate 3D reconstructions without a significant accuracy trade-off. We propose a novel network architecture and a training method that decouple the structure learning problem from the geometry reconstruction task. Our main intuition is that structure understanding is only necessary at the training stage and thus, by decoupling the two problems, we can obtain an accurate and efficient 3D point cloud reconstructions. Unlike previous methods, the approach we propose in this paper does not rely on computationally expensive operations and can be used to reconstruct, from monocular RGB images, accurate point clouds and meshes of arbitrary topological structures.

3. Method

We focus on reconstructing, in an efficient manner, accurate 3D point clouds from monocular RGB images. Point clouds are a light representation that can capture fine geometric details and can represent 3D shapes of arbitrary topologies. However, they are not structured and thus are difficult to process using modern deep neural networks. We address this problem using a novel Teacher-Student architecture where the teacher structures the data while the student maps input RGB images to the structured point clouds.

Intuitively, given a complex task, the Student does not need to fully solve the task from scratch. Instead, the Student only needs to understand the solution to the task. The Teacher, being more informed about the task, can explain (through supervision) the solution to the Student in a comprehensible manner. From the neural network perspective, as a Student, solving the unstructured nature of the point cloud representation to reconstruct point clouds from single RGB images can be seen as unnecessary. Computational power is often wasted on trying to solve this problem from unstructured data. The task of mapping images to 3D point clouds can be drastically simplified by first structuring the target point cloud. To this end, we present a Teacher network that learns to reconstruct point clouds based on its own perception of the input point cloud. We then train a Student network to understand the mapping between the input images and the newly annotated point clouds. Finally, we train a refinement network to further refine the generated point clouds.

3.1. The Teacher Network

The aim of the Teacher network, whose architecture is shown in Fig. 1, is to re-annotate and encode the ground-truth point clouds into latent representations that are easily interpretable by the Student network. It receives as input a point cloud, randomly sampled from the surface of the ground-truth 3D object, with no notion of structure. It outputs a latent representation and a 3D reconstruction of the input point cloud. In contrast to real-world points that are unstructured, points that are generated by the proposed neural network can be structured if the network is well-regularized at the training stage. Those structured points should be easier to learn and understand by another neural network.

Yang *et al.* [YFST18] introduced FoldingNet, an autoencoder that uses 2D grid deformations to constrain point cloud reconstruction. It uses 2D to 3D folding operations to map a set of 2D points to a 3D surface in a similar way to plane deformation. The deformed 3D points share the same structure as the input 2D grid. We propose a Teacher network that leverages the success of FoldingNet [YFST18] to generate structured point clouds that are easier to learn for the Student network. We adopt PointNet [QSMG17] as a global feature extractor module for unordered points. We extract a vector of dimension 1024 and pass it to a cascade of two fully-connected layers of 512 units each. We use ReLU as a non-linear activation for the hidden layer and quantization [OVK17] for the last layer. We adopt folding operations [YFST18] to decode the latent representation into 3D point clouds.

A folding decoder [YFST18] deforms a fixed initial 2D grid of points and has proven to be efficient in unsupervised semantic segmentation. We use a grid of 45×45 points, evenly spaced on a square of dimensions $[-0.3, 0.3]^2$. The original implementation of the folding operation [YFST18] uses fully-connected layers. We found that a deconvolution-based folding operation performs better than the original implementation. We implement it in the form of a decoder that contains two folding operations. The first one maps a fixed 2D grid to the 3D space using a cascade of three deconvolutional networks. The output spaces of the deconvolution layers are 512, 512, and 3. The second folding operation maps the output of the first folding operation to the point cloud. Each folding operation receives as input a concatenation of the latent vector with the input that needs to be deformed, *i.e.*, the 2D grid for the first folding operation and the output of the first folding operation for the second one.

3.2. Student Network

The Student network maps an input RGB image of size 224×224 to a 512D latent vector generated by the Teacher network. We adopt VGG19 [SZ14] with batch normalization as a feature extractor. We apply average pooling on the extracted features to obtain a $7 \times 7 \times 512$ latent representation, which is then processed using a cascade of three fully-connected layers. The first hidden layer maps the flattened feature vector into a vector of size 1024. The second hidden layer has 1024 units. Finally, a fully connected layer maps the generated features into the 512 latent vector that was generated by the Teacher network. We use ReLU activation for all the hidden layers. To generate the point cloud, we use a decoder that has

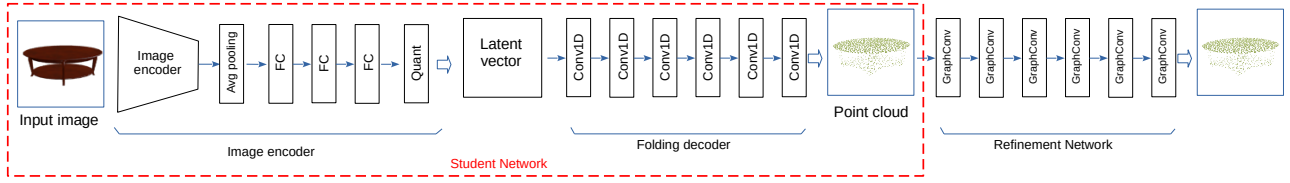


Figure 1: Overview of the proposed Teacher and Student network architecture. The former takes unstructured groundtruth points and outputs a structured point set, which is then used as a groundtruth label to train the Student network. Only the latter is used at test time.

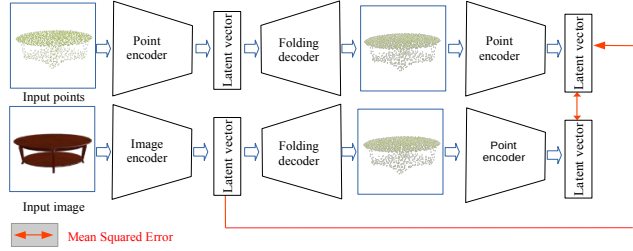


Figure 2: Overview of the training method. We use the point cloud generated by the Teacher network and a novel loss term, termed the Latent Distance, to train the Student network.

the same architecture as the decoder of the Teacher network; see Section 3.1.

3.3. Refinement Network

The Student network is trained using the annotations of the Teacher network. Thus, its performance is bounded by the performance of the Teacher network. To overcome this limitation, we propose a refinement network to refine the points generated by the Student network. These are unordered points that have a local structure within them. There are multiple ways to encode the local structure within the point clouds. For instance, Qi *et al.* [QSMG17, QYSG17] use a shared Multi-Layer Perceptron (MLP) across all points. More recent works [WSL*19, XDZQ21, NCKL19] use a graph obtained with the K-nearest neighbor algorithm. In this work, we propose to use the Alpha complex algorithm [EKS83] since it reconstructs surfaces and enables mesh-based operations such as surface pooling. Our refinement network receives a graph input and outputs a set of refined points. It is composed of six graph convolution layers having 16, 64, 128, 64, 16, 3 kernels, respectively, and a surface pooling layer that extracts surface points from the refined graph.

4. Loss Function

Point cloud accuracy assessment is one of the most fundamental problems in 3D reconstruction. The dominant trend in data-driven shape reconstruction is to either propose a better architecture, a better shape representation, or a better triangulation process. One research aspect that received less attention is the design of a better loss term that addresses the limitations of the widely used Chamfer Distance (CD). For two sets of point \mathcal{P} and \mathcal{Q} , the Chamfer Distance is given by

$$CD(\mathcal{P}, \mathcal{Q}) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \min_{q \in \mathcal{Q}} \|p - q\|_2^2 + \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \min_{p \in \mathcal{P}} \|p - q\|_2^2, \quad (1)$$

One of the limitations of the CD is the weak correlation between minimizing the CD and the visual quality of the reconstruction [WZL*18, GMJ19]. In fact, the CD is insensitive to subtle deformations; see Fig. 3. The mean CD error during the training phase on ShapeNet [CFG*15] is around 3×10^{-4} . Once the CD falls below the 3×10^{-4} threshold, it becomes insensitive to subtle deformations and gets saturated. Alternative measures include the Earth Mover Distance (EMD) [RTG00], which is computationally expensive since it solves the assignment optimization problem and relies on finding a unique bijection between point sets [QSMG17]. We thus propose to use as a loss function the difference between deep features, which is proven to be very efficient in assessing the perceptual similarity between images [ZIE*18]. Also, we observe that deep features capture high-level concepts that cannot be captured with low-level measures such as the CD or the EMD.

Let F be the encoder of the Teacher network. Then, the proposed Latent Distance (LD) is given by

$$LD(\mathcal{P}, \mathcal{Q}) = \frac{1}{|F(\mathcal{P})|} \|F(\mathcal{P}) - F(\mathcal{Q})\|_2^2. \quad (2)$$

Fig. 3 shows that, unlike the CD, the Latent Distance can effectively capture more discriminative information. In particular, the latent distance is large when the reconstruction is noisy. Thus, it penalizes noisy reconstructions and favors smooth ones. This is not the case with the CD.

Similar to the CD, the proposed Latent Distance is invariant to the way the input points are ordered. However, unlike the CD, the Student network does not need to solve the unstructured aspect of the point-cloud representation. For this reason, we seek to transfer the structure information recovered by the Teacher network to the Student network. We do this by matching the latent space of the Teacher network to the latent space of the Student network, using the Mean Square Error (MSE). The loss term of the Student network is then defined as:

$$\text{Loss} = LD(\mathcal{P}_{\text{gt}}, \mathcal{P}_{\text{rec}}) + \text{MSE}(L_{\text{Student}}(I), F(\mathcal{P}_{\text{gt}})), \quad (3)$$

where $L_{\text{Student}}(I)$ and $F(\mathcal{P}_{\text{gt}})$ are the latent vectors of the Student and the Teacher networks, respectively, \mathcal{P}_{gt} is the re-annotated ground truth point cloud, and \mathcal{P}_{rec} is the point cloud reconstructed by the Student network.

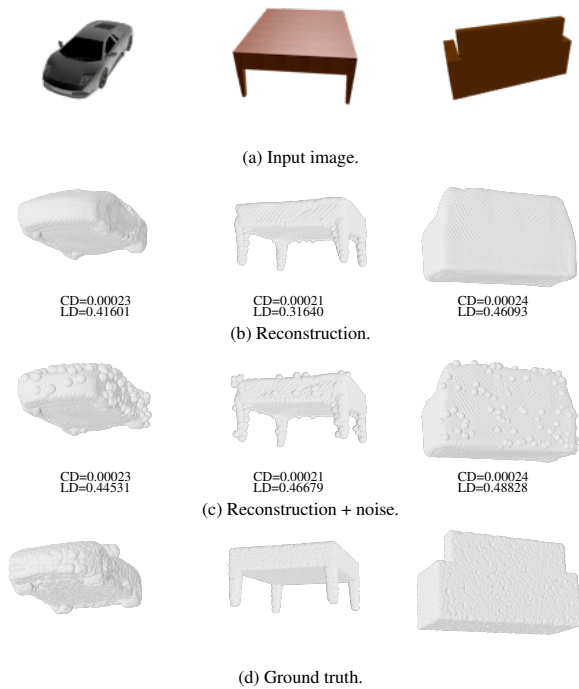


Figure 3: Sensitivity to subtle deformation: examples of reconstructions with and without added deformation. For all of the examples, the CD remains the same despite adding noise while the Latent Distance varies significantly.

The CD does not establish a one-to-one mapping between the predicted and the target 3D points. This makes it insensitive to subtle deformation as shown in Fig. 3. The Student network relies, during the training phase, on the annotations of the Teacher network, which was trained using the CD. Thus, the Student network inherits the CD limitation from the Teacher network. To overcome this, we train the refinement network using an adversarial loss. Our adversarial network, *i.e.*, the discriminator, has a similar architecture to our proposed point encoder network (Section 3.1). The only difference is in the last layer in which the fully-connected layer has only one unit and is followed by a Sigmoid activation. As the discriminator network is trained to only differentiate between ground-truth point clouds and the reconstructed point clouds, we added the CD to the loss term to ensure that the refinement network is reconstructing the correct object. The full loss term of the refinement stage is:

$$\text{Loss}_{ref} = \text{BCE}(\text{Disc}(\text{Rf}(\mathcal{P}_{rec})), \mathcal{Y}_{real}) + 1000 \times \text{CD}(\mathcal{P}_{gt}, \mathcal{P}_{rec}). \quad (4)$$

Here, BCE is the Binary Cross Entropy function, Rf the refinement network, \mathcal{Y}_{real} the annotations used to train the adversarial network, and Disc the discriminator network. We multiply the Chamfer distance by a factor of 1K to align it with the magnitude of the BCE and avoid mode collapse.

Table 1: We report, using the Chamfer Distance, the reconstruction error on 13 categories from the ShapeNet benchmark [CFG*15]. We use 3D-LMNet [MMAB18] and AttentionDPCR [LXL*19]’s evaluation protocols. The lower the value the better the performance.

Category	3D-LMNet Protocol				AttentionDPCR Protocol			
	PSGN	3D-LMNet	Ours	Ours(ICP)	PSG-FC	DensePCR	AttentionDPCR	Ours
Airplane	0.374	0.334	0.173	0.090	3.446	3.093	4.232	0.732
Bench	0.463	0.455	0.164	0.120	3.092	1.997	1.223	0.622
Cabinet	0.698	0.609	0.270	0.228	0.778	0.909	0.738	0.581
Car	0.520	0.455	0.163	0.139	0.925	1.069	0.931	0.485
Chair	0.639	0.641	0.339	0.243	3.591	2.630	1.507	0.991
Lamp	0.633	0.710	0.499	0.252	5.246	4.359	3.465	2.172
Monitor	0.615	0.640	0.328	0.215	1.748	1.629	1.339	1.033
Firearm	0.291	0.275	0.134	0.051	1.450	1.048	1.443	0.873
Couch	0.698	0.585	0.258	0.201	1.491	1.745	1.085	0.894
Speaker	0.875	0.810	0.434	0.336	1.251	1.398	1.111	1.169
Table	6.00	6.05	0.256	0.203	3.559	1.695	1.266	0.665
Telephone	0.456	0.463	0.181	0.119	1.157	0.925	0.881	0.613
Watercraft	0.438	0.437	0.202	0.115	1.327	1.601	1.603	1.043
Mean	0.562	0.540	0.262	0.178	2.236	1.847	1.603	0.913

5. Experiments

We evaluate the performance of the proposed framework on three different datasets: (1) ShapeNet [CFG*15] (Section 5.1), (2) Pix3D [SWZ*18] (Section 5.3), and (3) in-the-wild images (Section 5.4). For a fair comparison on the ShapeNet benchmark [CFG*15], we train and evaluate the proposed framework on a subset of 13 categories. We adopt the evaluation protocols previously used by the state-of-the-art methods. These are detailed in the Supplementary Material. Finally, we perform an ablation study to assess the contribution of the different components of the proposed approach (See Section 5 in the Supplementary Material).

5.1. Performance on ShapeNet

Following the protocols described in [WZL*18, GMJ19], we train and test our method on a subset containing 13 categories of ShapeNet [CFG*15]. For a fair evaluation, we use the images rendered by Choy *et al.* [CXG*16], which we split into training, validation, and test following the training protocol described in [GMJ19].

5.1.1. Quantitative Evaluation

Table 1 evaluates, on the ShapeNet benchmark [CFG*15], the performance of our method and compares it to other point cloud-based approaches such as PSGN-FC [FSG17], DensePCR [MR19], and AttentionDPCR [LXL*19]. From this table, we can see that our method outperforms the state-of-the-art in 12 out of the 13 shape categories, with a performance increase of more than 97%. Our method requires 6.33 ms to generate a point cloud of 10K points on a machine equipped with an i9 processor and a Nvidia Titan RTX GPU card. This is very efficient compared to the state-of-the-art, *e.g.*, Mesh R-CNN [GMJ19], which requires 679ms and DefTet [GCX*20], which requires 200 ms. PSGN [FSG17], on the other hand, requires only 3.61ms but is 61% less accurate than ours as shown in Table 2.

The Student network has been trained using the annotations produced by the Teacher. Thus, the reconstructed point clouds follow a probability distribution that is slightly different from the ground-truth distribution. This results in slightly larger Chamfer

Table 2: Evaluation on ShapeNet dataset following Pixel2Mesh protocol. We report the CD and F1 scores. * refers to the results where we sample 40k predicted points and 16384 ground truth points to align with PCDNet(UpResGraphX) [NCKL19] and Depth Intermediation [ZKG18].

Model	CD ↓	F1 ↑		Camera intrinsic
		τ	2τ	
N3MR [KUH18]	2.629	33.80	47.72	-
3D-R2N2 [CXG*16]	1.445	39.01	54.62	-
PSGN [FSG17]	0.593	48.58	69.78	×
Pixel2Mesh [WZL*18]	0.591	59.71	74.19	✓
MVD [SFM18]	-	66.39	-	-
TopoNet [BCTCL22]	0.500	66.01	78.20	✓
GEOmetrics [SFRM19]	-	67.37	-	✓
BSPNet [CTZ20]	0.465	-	-	×
Mesh R-CNN (Best) [GMJ19]	0.306	74.84	85.75	✓
Ours (No refinement)	0.310	74.4	85.2	×
Ours	0.255	79.5	88.7	×
PCDNet* (UpResGraphX) [NCKL19]	0.252	-	-	✓
Depth Intermediation* [ZKG18]	0.246	-	-	✓
Ours*	0.231	82.0	89.7	×
Teacher	0.097	89.26	95.40	×

Distance values. To alleviate this issue and ensure a fair comparison with the state of the art, we sample surface-extracted point clouds after applying the refinement stage. Table 2 compares the performance of our method to PCDNet(UpResGraphX) [NCKL19] and Depth Intermediation [ZKG18] following the evaluation protocol of Pixel2Mesh [WZL*18]. Depth Intermediation [ZKG18] uses 3D-CNN to deform a 3D grid. PCDNet(UpResGraphX) [NCKL19] uses a novel graph-x operation that operates similarly to graph convolution. Both methods assume a known camera intrinsic matrix. This limits their application in real-world scenarios. Despite the fact that our network does not rely on known camera intrinsics to reconstruct 3D objects, it outperforms the state-of-the-art methods in terms of accuracy. This is mainly due to our improved training methodology.

Table 1 compares the performance of our proposed method to 3D-LMNet [MMAB18] and PSGN [FSG17] using the evaluation protocol proposed by 3D-LMNet [MMAB18]. In contrast to the proposed method, 3D-LMNet [MMAB18] relies on a simple latent matching without structure learning. As seen in Table 1, the proposed method outperforms 3D-LMNet [MMAB18] by a large margin. From this observation, we conclude that learning structured points is more efficient than non-structured ones. Additionally, MandiKal *et al.* [MMAB18] rely on the Iterative Closest Point (ICP) algorithm to align the reconstructed point clouds with the ground truth points, which further enhances the reconstruction accuracy. We report the performance of the proposed method before and after applying the ICP.

5.1.2. Qualitative Evaluation

Figure 4 compares the visual quality of our method to PSGN [FSG17] and 3D-LMNet [MMAB18]. PSGN [FSG17] is trained in an end-to-end manner using the Chamfer Distance as a loss function. 3D point clouds generated using 3D-

LMNet [MMAB18] are more accurate than points generated by PSGN [FSG17]. However, 3D-LMNet [MMAB18] fails to faithfully generate thin structures such as chair legs; see for example the fourth and fifth rows of Figure 4. The structure of the Teacher network can be observed on flat surfaces of our reconstructions. There, we can see that, compared to the ground-truth points, points generated by our method seem to have an organized pattern; see for example the first three rows of Figure 4.

Figure 6 illustrates the structure learned by the teacher network. We color the points in the initial grid based on their order. Then, we deform the grid using the Teacher network and use the same color mapping. As seen in Figure 6, the learned parametrization provides a plausible mapping (correspondence) between parts of Object A and Object B. For instance, both the front right legs of the tables seen in Figure 6, first row, are colored in light green while the back right legs are colored in orange.

5.2. Application to Mesh Reconstruction

The proposed refinement network generates a mesh. In this section, we compare the performance of our mesh reconstruction approach with the state-of-the-art. In addition to the point-based methods, we also consider (1) mesh-based methods such as N3MR [KUH18], Pixel2Mesh [WZL*18], and GEOmetrics [SFRM19], which are template based, and Mesh R-CNN [GMJ19], which reconstructs meshes of arbitrary topology but uses voxels as an intermediate representation, and (2) volumetric methods such as 3D-R2N2 [CXG*16] and MVD [SFM18]. We use the Pixel2Mesh [WZL*18] evaluation protocol. Table 2 reports the performance. Table 3 compares the performance of our proposed method to Mesh R-CNN [GMJ19] and Pixel2Mesh [WZL*18] following the evaluation protocol of Mesh R-CNN [GMJ19]. As illustrated in Table 3, our proposed method significantly outperforms Mesh R-CNN [GMJ19] and Pixel2Mesh [WZL*18] in terms of accuracy.

Figure 5 compares the visual aspect of our mesh reconstructions to Pixel2Mesh [WZL*18] and Mesh R-CNN [GMJ19]. Since Pixel2Mesh [WZL*18] deforms a sphere, it cannot reconstruct objects of complex topological structures. Mesh R-CNN [GMJ19] relies on low-resolution volumetric grids as an intermediate representation. This results in non-smooth surfaces. Our method is capable of reconstructing objects of arbitrary topological structures as shown in Figure 5. Unlike Pixel2Mesh [WZL*18] and Mesh R-CNN [GMJ19], our method was not trained to generate meshes and does not assume a known camera intrinsic matrix. Nevertheless, it generates reconstructions of a higher fidelity to the input images than mesh-based solutions. As shown in Figure 5, our method can reconstruct meshes of arbitrary topological structures more accurately than Mesh R-CNN [GMJ19]; see the rows 2, 3, and 4 of Figure 5.

5.3. Performance on Pix3D

We evaluate the performance of our proposed framework on Pix3D benchmark [SWZ*18], which is composed of 10,069 real images of 395 CAD models. Compared to ShapeNet [CFG*15], Pix3D is more challenging due to (1) the misalignment between the images

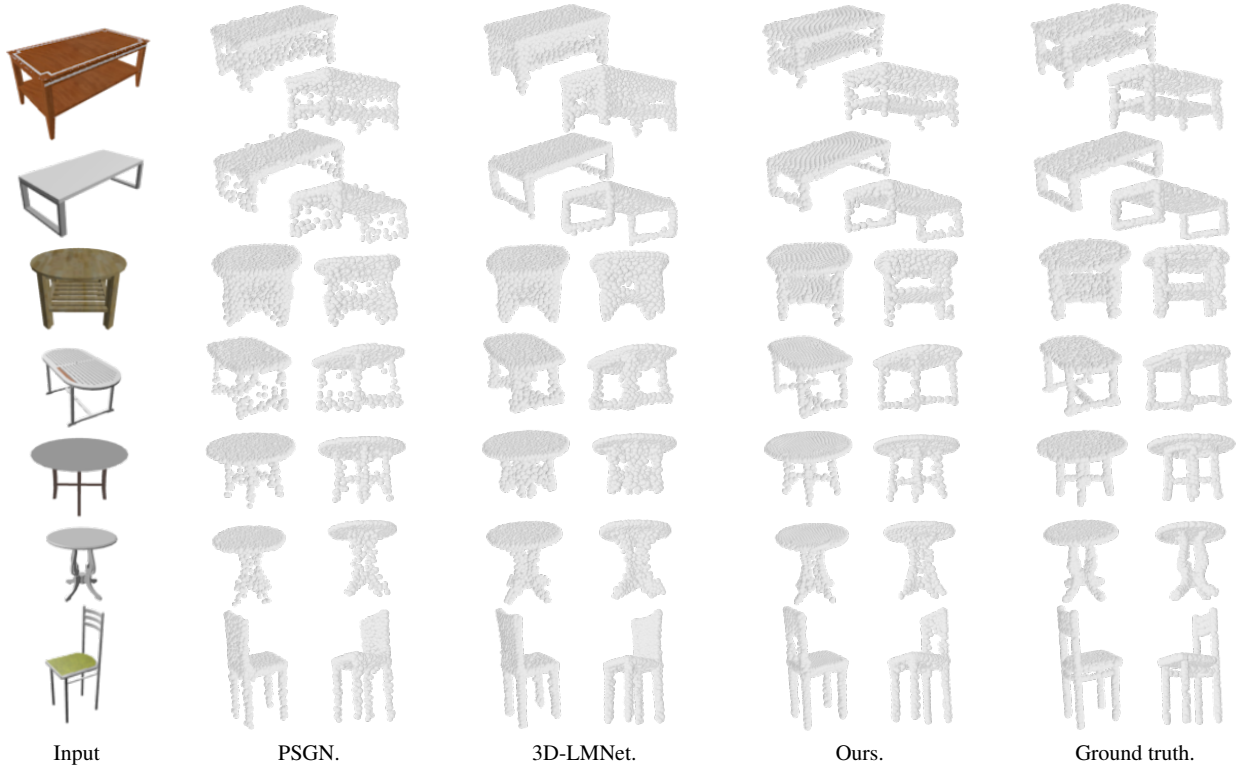


Figure 4: Qualitative comparison, using ShapeNet [CFG*15], of the visual aspect of our reconstructions to PSGN [FSG17] and 3D-LMNet [MMAB18].

Table 3: Reconstruction error on ShapeNet using the scale-invariant protocol of Mesh R-CNN. We compare to the state-of-the-art and to an ablated model of Mesh R-CNN.

	Full Test Set							Holes Test Set						
	CD ↓	Normal ↑	F ₁ ^{0.1} ↑	F ₁ ^{0.3} ↑	F ₁ ^{0.5} ↑	V	F	CD ↓	Normal ↑	F ₁ ^{0.1} ↑	F ₁ ^{0.3} ↑	F ₁ ^{0.5} ↑	V	F
Pixel2Mesh	0.265	0.729	29.9	76.2	89.0	2466 ± 0	4928 ± 0	0.273	0.733	30.8	76.5	88.9	2466 ± 0	4928 ± 0
Mesh R-CNN (Best)	0.133	0.729	38.8	86.8	95.1	1899 ± 928	3800 ± 1861	0.130	0.725	41.7	86.7	94.9	2291 ± 903	4595 ± 1814
Mesh R-CNN (Pretty)	0.171	0.713	35.1	82.6	93.2	1896 ± 928	3795 ± 1861	0.171	0.700	37.1	82.4	92.7	2292 ± 902	4598 ± 1812
Ours	0.108	0.611	44.5	89.2	96.3	3161 ± 569	31311 ± 7455	0.108	0.588	46.5	88.9	96.0	3165 ± 576	30887 ± 6537

and their corresponding 3D objects, (2) the presence of occlusions, (3) the different lighting conditions, (4) the different camera intrinsic matrices, and (5) the limited size of the dataset. We compare our performance to Mesh R-CNN [GMJ19] following the evaluation protocol of Mesh R-CNN; see Table 4. The proposed method has a reconstruction error (Chamfer Distance) that is 21% lower than Mesh R-CNN [GMJ19]. However, Mesh R-CNN [GMJ19] reports higher F₁ scores than the proposed method. In fact, unlike ours, Mesh R-CNN [GMJ19] assumes a known camera intrinsic matrix which results in better alignment between the reconstruction and the input image. The proposed method achieves comparable performance without relying on such information. This makes the proposed method applicable in a wider range of applications where the camera’s intrinsic information cannot be easily obtained.

Table 4: Quantitative evaluation on Pix3D. We report the CD and the F₁ scores of Mesh R-CNN and the proposed model.

Model	CD ↓	F ₁ ^{0.1} ↑	F ₁ ^{0.3} ↑	F ₁ ^{0.5} ↑
Mesh R-CNN	1.11	18.7	56.4	73.5
Ours	0.87	16.6	55.4	72.2

5.4. In-The-Wild Evaluation

We test our model, which was trained on the synthetic dataset of ShapeNet [CFG*15], on in-the-wild images from the Internet. Figure 7 qualitatively compares its performance with PSGN [FSG17] and 3D-LMNet [MMAB18]. 3D-LMNet [MMAB18] overfits its training set and fails to generalize to real-world images. Our model, which was trained using synthetic data, generalizes well to unseen

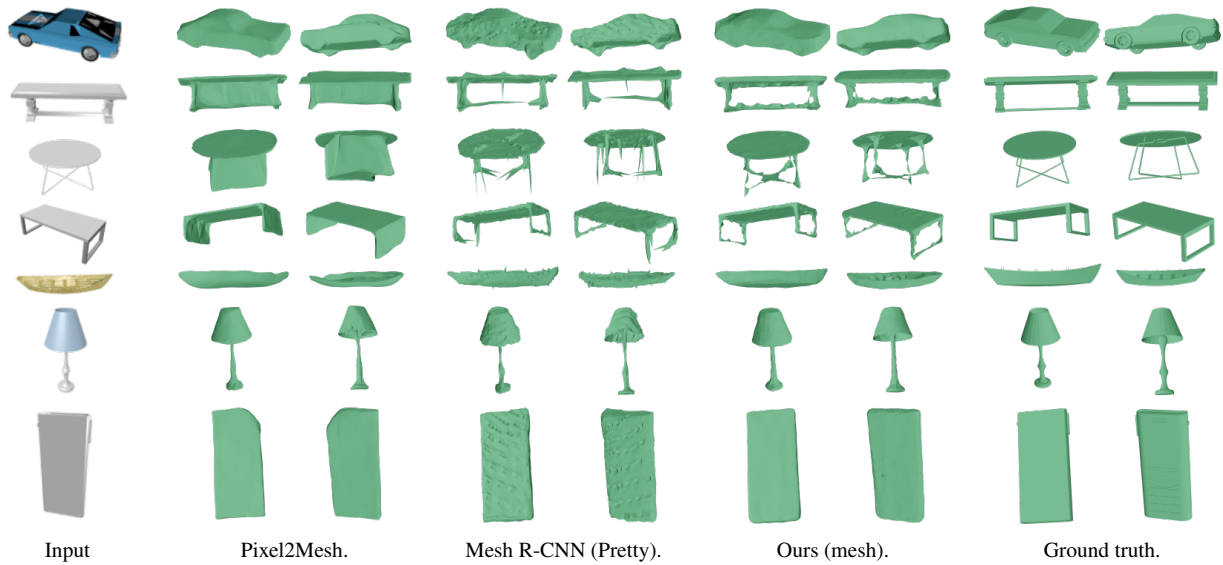


Figure 5: Qualitative comparison, using ShapeNet [CFG*15], of the proposed mesh-based 3D reconstruction to Pixel2Mesh [WZL*18] and Mesh R-CNN [GMJ19].

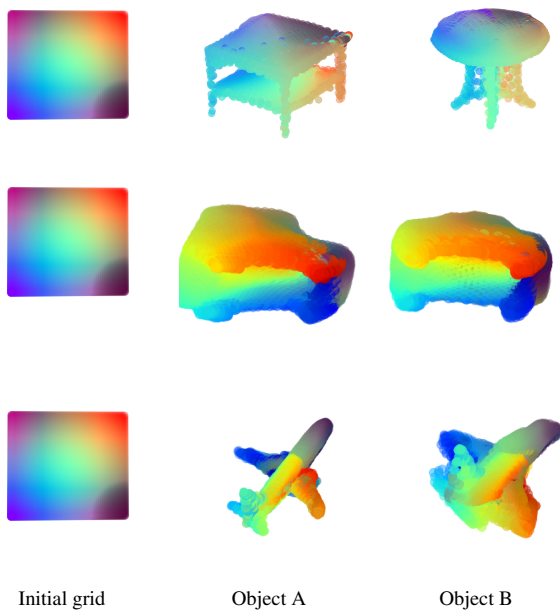


Figure 6: Parametrization results obtained by the Teacher network: We observe that the generated parametrization provides a direct mapping (correspondence) between Object A and Object B.

real-world images and generates reconstructions of a higher fidelity to the input images than PSGN [FSG17].

6. Conclusion

We proposed in this paper a novel framework for point cloud reconstruction from single RGB images. Our novel training methodology decouples the structure-learning problem from the reconstruction problem by relying on two different networks: a **Teacher** network and a **Student** network. Such methodology allows the reconstruction of accurate point clouds without relying on computationally costly operations, *e.g.*, graph-X [NCKL19], nor assuming a known camera intrinsic matrix. Additionally, we proposed a novel loss to train the Student network. In contrast to the widely used Chamfer Distance that estimates the average error of individual points, our Latent distance relies on deep features to compute a global error between 3D shapes. We use the Chamfer Distance to evaluate the performance of this proposed model and to train the Teacher network which represents a limitation for the current method. As a future work, we will focus on fixing the limitations of this evaluation metric.

7. Acknowledgment

This research was supported partially by the Australian Government through the Australian Research Council’s Discovery Projects funding scheme (project DP220102197). The views expressed herein are those of the authors and are not necessarily those of the Australian Government or Australian Research Council.

References

[AB22] ARSHAD M. S., BEKSI W. J.: Automated reconstruction of 3d open surfaces from sparse point clouds. In *2022 IEEE International*

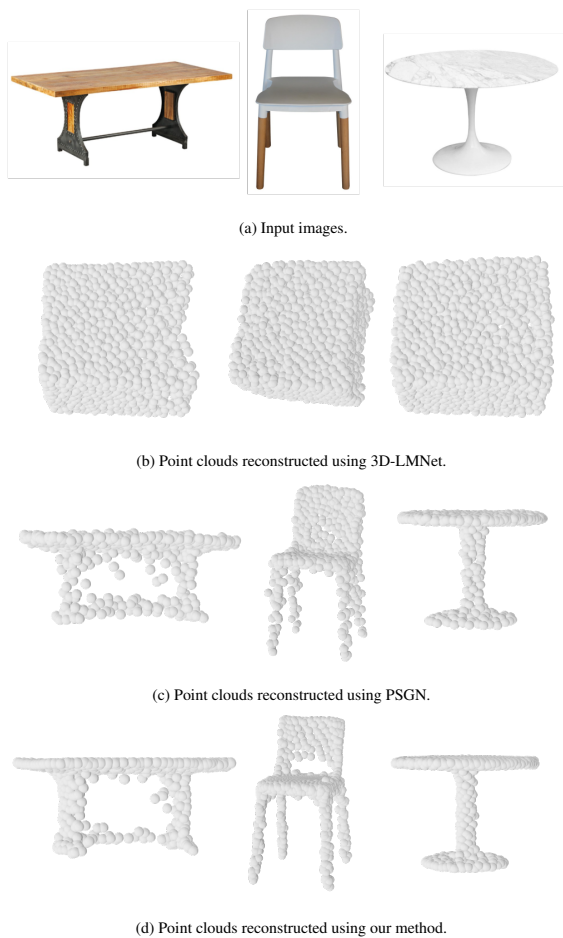


Figure 7: Reconstruction from real-world images. Our method generalizes well to unseen images despite being trained on ShapeNet [CFG*15], which is a synthetic dataset.

Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct) (2022), pp. 216–221. doi:10.1109/ISMAR-Adjunct57072.2022.00048. 1, 2

- [ASY22] ANISETTY S., SARAVANABAVAN V., YIYU C.: Learning to regulate 3d head shape by removing occluding hair from in-the-wild images. In *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)* (2022), pp. 403–408. doi: 10.1109/ISMAR-Adjunct57072.2022.00087. 1
- [BCTCL22] BEN CHARRADA T., TABIA H., CHETOUANI A., LAGA H.: Toponet: Topology learning for 3d reconstruction of objects of arbitrary genus. In *Computer Graphics Forum* (2022), vol. 41, Wiley Online Library, pp. 336–347. 1, 2, 6
- [BPM20] BOULCH A., PUY G., MARLET R.: Fkaconv: Feature-kernel alignment for point cloud convolution. In *Proceedings of the Asian Conference on Computer Vision* (2020). 3
- [CFG*15] 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 preprint arXiv:1512.03012* (2015). 4, 5, 6, 7, 8, 9
- [CLK*18] CAO Y.-P., LIU Z.-N., KUANG Z.-F., KOBBELT L., HU S.-M.: Learning to reconstruct high-quality 3D shapes with cascaded fully convolutional networks. In *ECCV* (2018). 2

- [CSO*18] CHERABIER I., SCHONBERGER J. L., OSWALD M. R., POLLEFEYS M., GEIGER A.: Learning Priors for Semantic 3D Reconstruction. In *ECCV* (2018). 2
- [CTZ20] CHEN Z., TAGLIASACCHI A., ZHANG H.: Bsp-net: Generating compact meshes via binary space partitioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 45–54. 6
- [CXG*16] CHOY C. B., XU D., GWAK J., CHEN K., SAVARESE S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision* (2016), Springer, pp. 628–644. 1, 5, 6
- [DP22] DUGGAL S., PATHAK D.: Topologically-aware deformation fields for single-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 1536–1546. 2
- [DRQN17] DAI A., RUIZHONGTAI QI C., NIESSNER M.: Shape completion using 3D-encoder-predictor CNNs and shape synthesis. In *IEEE CVPR* (2017), pp. 5868–5877. 2
- [EKS83] EDELSBRUNNER H., KIRKPATRICK D., SEIDEL R.: On the shape of a set of points in the plane. *IEEE Transactions on information theory* 29, 4 (1983), 551–559. 4
- [FSG17] FAN H., SU H., GUIBAS L.: A point set generation network for 3D object reconstruction from a single image. In *IEEE CVPR* (2017), vol. 38. 1, 3, 5, 6, 7, 8
- [GCX*20] GAO J., CHEN W., XIANG T., JACOBSON A., MCGUIRE M., FIDLER S.: Learning deformable tetrahedral meshes for 3d reconstruction. *Advances In Neural Information Processing Systems* 33 (2020), 9936–9947. 5
- [GMJ19] GKIOXARI G., MALIK J., JOHNSON J.: Mesh r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 9785–9795. 2, 3, 4, 5, 6, 7, 8
- [GWM18] GADELHA M., WANG R., MAJI S.: Multiresolution tree networks for 3D point cloud processing. In *ECCV* (2018), pp. 103–118. 1
- [HLB21] HAN X., LAGA H., BENNAMOUN M.: Image-based 3d object reconstruction: State-of-the-art and trends in the deep learning era. *IEEE transactions on pattern analysis and machine intelligence* (2021). 1, 2
- [HTM19] HANE C., TULSIANI S., MALIK J.: Hierarchical Surface Prediction. *IEEE PAMI*, 1 (2019), 1–1. 2
- [HWX*21] HU T., WANG L., XU X., LIU S., JIA J.: Self-supervised 3d mesh reconstruction from single images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 6002–6011. 2
- [JSQJ18] JIANG L., SHI S., QI X., JIA J.: GAL: Geometric Adversarial Loss for Single-View 3D-Object Reconstruction. In *ECCV* (2018). 1
- [KLR18] KUNDU A., LI Y., REHG J. M.: 3D-RCNN: Instance-Level 3D Object Reconstruction via Render-and-Compare. In *IEEE CVPR* (2018), pp. 3559–3568. 2
- [KTEM18] KANAZAWA A., TULSIANI S., EFROS A. A., MALIK J.: Learning Category-Specific Mesh Reconstruction from Image Collections. *ECCV* (2018). 1
- [KUH18] KATO H., USHIKU Y., HARADA T.: Neural 3D Mesh Renderer. In *IEEE CVPR* (2018). 1, 6
- [LBS*18] LI Y., BU R., SUN M., WU W., DI X., CHEN B.: Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems* 31 (2018). 3
- [LCRCN22] LI L., CARNELL S., REINERS D., CRUZ-NEIRA C.: A system design to create mixed 360 video and 3d content for virtual field trip. In *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)* (2022), pp. 198–201. doi: 10.1109/ISMAR-Adjunct57072.2022.00044. 1

- [LFB*23] LIU Z., FENG Y., BLACK M. J., NOWROUZEZAHRAI D., PAULL L., LIU W.: Meshdiffusion: Score-based generative 3d mesh modeling. *arXiv preprint arXiv:2303.08133* (2023). 2
- [LGCR19] LI K., GARG R., CAI M., REID I.: Single-view object shape reconstruction using deep shape prior and silhouette. *arXiv:1811.11921* (2019). 1
- [LGOA18] LIU S., GILES C. L., ORORBIA I., ALEXANDER G.: Learning a Hierarchical Latent-Variable Model of 3D Shapes. *International Conference on 3D Vision* (2018). 2
- [LJBB20] LAGA H., JOSPIN L. V., BOUSSAID F., BENNAMOUN M.: A survey on deep learning techniques for stereo-based depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020). 2
- [LKL18] LIN C.-H., KONG C., LUCEY S.: Learning Efficient Point Cloud Generation for Dense 3D Object Reconstruction. *AAAI* (2018). 1
- [LPZR18] LI K., PHAM T., ZHAN H., REID I.: Efficient dense point cloud object reconstruction using deformation vector fields. In *ECCV* (2018), pp. 497–513. 1
- [LWL20] LIN C.-H., WANG C., LUCEY S.: Sdf-srn: Learning signed distance 3d object reconstruction from static images. *Advances in Neural Information Processing Systems 33* (2020), 11453–11464. 2
- [LXC*17] LI J., XU K., CHAUDHURI S., YUMER E., ZHANG H., GUIBAS L.: GRASS: Generative Recursive Autoencoders for Shape Structures. *ACM TOG* 36, 4 (2017), 52. 2
- [LXL*19] LU Q., XIAO M., LU Y., YUAN X., YU Y.: Attention-based dense point cloud reconstruction from a single image. *IEEE Access* 7 (2019), 137420–137431. 5
- [LZ21] LI M., ZHANG H.: D2im-net: Learning detail disentangled implicit fields from single images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 10246–10255. 2
- [LZZ*19] LI C.-L., ZAHEER M., ZHANG Y., POCZOS B., SALAKHUTDINOV R.: Point cloud GAN. *ICLR Workshop on Deep Generative Models for Highly Structured Data* (2019). 1
- [MBM*17] MONTI F., BOSCAINI D., MASCI J., RODOLA E., SVOBODA J., BRONSTEIN M. M.: Geometric deep learning on graphs and manifolds using mixture model cnns. In *CVPR* (2017), vol. 1, p. 3. 2
- [MMAB18] MANDIKAL P., MURTHY N., AGARWAL M., BABU R. V.: 3D-LMNet: Latent Embedding Matching for Accurate and Diverse 3D Point Cloud Reconstruction from a Single Image. *BMVC* (2018), 662–674. 1, 5, 6, 7
- [MON*19] MESCHEDER L., OECHSLE M., NIEMEYER M., NOWOZIN S., GEIGER A.: Occupancy Networks: Learning 3D Reconstruction in Function Space. *IEEE CVPR* (2019). 2
- [MR19] MANDIKAL P., RADHAKRISHNAN V. B.: Dense 3D Point Cloud Reconstruction Using a Deep Pyramid Network. In *IEEE WACV* (2019), pp. 1052–1060. 1, 5
- [NCKL19] NGUYEN A.-D., CHOI S., KIM W., LEE S.: Graphx-convolution for point cloud deformation in 2d-to-3d conversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 8628–8637. 4, 6, 8
- [NLX18] NIU C., LI J., XU K.: Im2Struct: Recovering 3D Shape Structure from a Single RGB Image. *IEEE CVPR 4096* (2018), 80. 2
- [OVK17] OORD A. V. D., VINIYALS O., KAVUKCUOGLU K.: Neural discrete representation learning. *arXiv preprint arXiv:1711.00937* (2017). 3
- [PEW21] PING G., ESFAHANI M. A., WANG H.: Visual enhanced 3d point cloud reconstruction from a single image. *arXiv preprint arXiv:2108.07685* (2021). 1
- [PFS*19] PARK J. J., FLORENCE P., STRAUB J., NEWCOMBE R., LOVEGROVE S.: DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *IEEE CVPR* (2019), pp. 165–174. 2
- [PHC*19] PAN J., HAN X., CHEN W., TANG J., JIA K.: Deep mesh reconstruction from single rgb images via topology modification networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 9964–9973. 1, 2
- [PHTPS20] PEÑA-TAPIA E., HACHIUMA R., PASQUALI A., SAITO H.: Lcr-smpl: Toward real-time human detection and 3d reconstruction from a single rgb image. In *2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)* (2020), pp. 211–212. doi:10.1109/ISMAR-Adjunct51615.2020.00062. 1
- [QSMG17] QI C. R., SU H., MO K., GUIBAS L. J.: PointNet: Deep learning on point sets for 3d classification and segmentation. In *IEEE CVPR* (2017), pp. 652–660. 3, 4
- [QYSG17] QI C. R., YI L., SU H., GUIBAS L. J.: PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *NIPS* (2017), pp. 5099–5108. 3, 4
- [RTG00] RUBNER Y., TOMASI C., GUIBAS L. J.: The earth mover's distance as a metric for image retrieval. *International journal of computer vision* 40, 2 (2000), 99–121. 4
- [RUG17] RIEGLER G., ULUSOY A. O., GEIGER A.: OctNet: Learning deep 3D representations at high resolutions. In *IEEE CVPR* (2017), vol. 3. 2
- [SFM18] SMITH E., FUJIMOTO S., MEGER D.: Multi-view silhouette and depth decomposition for high resolution 3d object representation. In *Advances in Neural Information Processing Systems* (2018), pp. 6478–6488. 6
- [SFRM19] SMITH E. J., FUJIMOTO S., ROMERO A., MEGER D.: Geometrics: Exploiting geometric structure for graph-encoded objects. *arXiv preprint arXiv:1901.11461* (2019). 6
- [SKKP22] STEDMAN H., KOCER B. B., KOVAC M., PAWAR V. M.: Vrtab-map: A configurable immersive teleoperation framework with on-line 3d reconstruction. In *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)* (2022), pp. 104–110. doi:10.1109/ISMAR-Adjunct57072.2022.00029. 1
- [SWL*18] SUN Y., WANG Y., LIU Z., SIEGEL J. E., SARMA S. E.: PointGrow: Autoregressively learned point cloud generation with self-attention. *arXiv:1810.05591* (2018). 1
- [SWZ*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* (2018), pp. 2974–2983. 1, 5, 6
- [SZ14] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014). 3
- [Tan22] TANAKA K.: 3d scene reconstruction from monocular spherical video with motion parallax. In *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)* (2022), pp. 191–197. doi:10.1109/ISMAR-Adjunct57072.2022.00043. 1
- [TDB17] TATARCHENKO M., DOSOVITSKIY A., BROX T.: Octree generating networks: Efficient convolutional architectures for high-resolution 3D outputs. In *IEEE CVPR* (2017), pp. 2088–2096. 2
- [TEM18] TULSIANI S., EFROS A. A., MALIK J.: Multi-View Consistency as Supervisory Signal for Learning Shape and Pose Prediction. In *IEEE CVPR* (2018). 2
- [THP*19] TANG J., HAN X., PAN J., JIA K., TONG X.: A skeleton-bridged deep learning approach for generating meshes of complex topologies from single rgb images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 4541–4550. 2, 3

- [TQD*19] THOMAS H., QI C. R., DESCHAUD J.-E., MARCOTEGUI B., GOULETTE F., GUIBAS L. J.: Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 6411–6420. 3
- [TZEM17] TULSIANI S., ZHOU T., EFROS A. A., MALIK J.: Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *IEEE CVPR* (2017), vol. 1, p. 3. 1
- [WCL19] WU Y.-C., CHAN L., LIN W.-C.: Tangible and visible 3d object reconstruction in augmented reality. In *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (2019), pp. 26–36. doi:10.1109/ISMAR.2019.00–30. 1
- [WLG*17] WANG P.-S., LIU Y., GUO Y.-X., SUN C.-Y., TONG X.: O-CNN: Octree-based convolutional neural networks for 3D shape analysis. *ACM TOG* 36, 4 (2017), 72. 2
- [WQF19] WU W., QI Z., FUXIN L.: Pointconv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 9621–9630. 3
- [WSL18] WANG J., SUN B., LU Y.: MVPNet: Multi-View Point Regression Networks for 3D Object Reconstruction from A Single Image. *arXiv:1811.09410* (2018). 1
- [WSL*19] WANG Y., SUN Y., LIU Z., SARMA S. E., BRONSTEIN M. M., SOLOMON J. M.: Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)* 38, 5 (2019), 1–12. 4
- [WWX*17] WU J., WANG Y., XUE T., SUN X., FREEMAN B., TENENBAUM J.: MarrNet: 3D shape reconstruction via 2.5D sketches. In *NIPS* (2017), pp. 540–550. 1
- [WZC*22] WEN C., ZHANG Y., CAO C., LI Z., XUE X., FU Y.: Pixel2mesh++: 3d mesh generation and refinement from multi-view images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 2 (2022), 2166–2180. 2
- [WZL*18] WANG N., ZHANG Y., LI Z., FU Y., LIU W., JIANG Y.-G.: Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 52–67. 1, 2, 4, 5, 6, 8
- [XDZQ21] XU M., DING R., ZHAO H., QI X.: Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 3173–3182. 4
- [XFX*18] XU Y., FAN T., XU M., ZENG L., QIAO Y.: Spidercnn: Deep learning on point sets with parameterized convolutional filters. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 87–102. 3
- [YFST18] YANG Y., FENG C., SHEN Y., TIAN D.: Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 206–215. 2, 3
- [ZIE*18] ZHANG R., ISOLA P., EFROS A. A., SHECHTMAN E., WANG O.: The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 586–595. 4
- [ZKG18] ZENG W., KARAOGU S., GEVERS T.: Inferring Point Clouds from Single Monocular Images by Depth Intermediation. *arXiv:1812.01402* (2018). 1, 6