Topology-Aware Single-Image 3D Shape Reconstruction

Qimin Chen[1] Vincent Nguyen[1] Feng Han[2] Raimondas Kiveris[2] Zhuowen Tu[1]
{qic003,vvn012}@ucsd.edu {bladehan,rkiveris}@google.com ztu@ucsd.edu
[1]UC San Diego [2]Google Inc.

Abstract

We make an attempt to address topology-awareness for 3D shape reconstruction. Two types of high-level shape typologies are being studied here, namely genus (number of cuttings/holes) and connectivity (number of connected components), which are of great importance in 3D object reconstruction/understanding but have been thus far disjoint from the existing dense voxel-wise prediction literature. We propose a topology-aware shape autoencoder component (TPWCoder) by approximating topology property functions such as genus and connectivity with neural networks from the latent variables. TPWCoder can be directly combined with the existing 3D shape reconstruction pipelines for end-to-end training and prediction. On the challenging A Big CAD Model Dataset (ABC), TPWCoder demonstrates a noticeable quantitative and qualitative improvement over the competing methods, and it also shows improved quantitative result on the ShapeNet dataset.

1. Introduction

Large progress in computer vision has been made for the dense prediction tasks using end-to-end pixel-wise training in a range of applications such as semantic segmentation [32, 9], edge detection [62], depth estimation [31], image denoising [66], super-resolution [27], medical image segmentation [43, 19], and single-image 3D shape reconstruction [55]. These tasks have been carried out mostly in a bottom-up fashion through a forward process. The visual perception tasks have shown to engage as well top-down knowledge [2] that provides strong regularities about the scene layout, spatial configurations, and shape information [21]. Top-down knowledge has been primarily viewed as a prior in the non-deep-learning based algorithms [13, 68]. Attempts to explicitly introduce the top-down prior for semantic labeling exist [61] but with multiple limitations and requirements.

The literature of 3D reconstruction is very rich in photogrammetry [30, 11] and computer vision [20, 33] ranging from structure from motion [52], stereo [4], multiview [38, 29], depth-sensor based [36, 69], large-scale computing [1], single-image based [18], and hybrid approaches [39]. The 3D reconstruction tasks become particularly important and useful in the modern big data era.

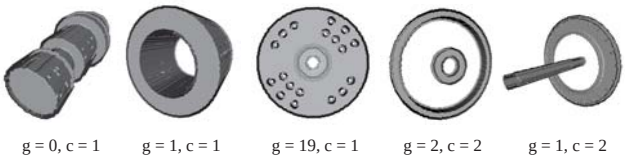

Figure 1: Some examples from the ABC dataset [28] with different number of genus g and connectivity c.

From a different angle, there has been a renewed interest for automatic single-view 3D reconstruction [55, 67, 49, 25] using convolutional neural networks for dense voxel-wise prediction. On one hand, the reconstruction results from 2D single-view image on the 3D ShapeNet [58] are impressive [55] and have seen a steady improvement [55, 57, 67] over the time; on the other hand, state-of-the-art methods [55, 63, 47] produce results that are still not satisfactory on challenging datasets such as the ABC dataset [28] in which objects exhibit a wide range of topological variations with a varying number of holes and connected components (see *e.g.* Figures 1 and 5). It is therefore important to move beyond the current pixel-wise/voxel-wise learning frameworks by jointly taking into account the voxel-wise reconstruction difference and the high-level geometric and topological properties for the 3D objects.

In this paper, we make an attempt to address topology-awareness for single-image 3D shape reconstruction by proposing topology-aware shape autoencoder (TPWCoder) that learns to approximate 3D topological functions. Here we emphasize the need of engaging differentiable structural and topological regularities that can be seamlessly combined with the current end-to-end deep learning frameworks. This step will help improve the current results for shape segmentation and reconstruction.

The main contributions of our work are listed below.

- Development of the **topology-aware** shape autoencoder (TPWCoder) to address high-level topological properties such as genus and connectivity for 3D shape reconstruction.
- Design of **differentiable** topological loss to combine with an end-to-end 3D reconstruction algorithm, MarrNet [55].
- **Noticeable** qualitative and quantitative improvement over the state-of-the-art on the challenging ABC [28] dataset.

2. Related work

Reconstructing the 3D shape of an object from its single 2D image is challenging due to the intrinsic ambiguity and the immensely large solution space for the problem. Owing to the rapid development of 3D objects synthetic datasets such as ShapeNet [6], ABC [28], effective algorithms have been developed to approach objects reconstruction in voxel as well as in point clouds [16] and octave trees [42, 48].

Single-image 3D Reconstruction. Early attempts using non-deep learning based methods in 3D shape reconstruction have been extensively studied. Huang *et al*. [24] presented an assembly-based method to reconstruct 3D shape by means of composing parts from existing 3D models. This approach, however, strongly depends on the availability of initial segmentation of the existing 3D shapes. Thanks to the significant progress made in deep learning, researchers have built more neural networks [26, 50, 55, 57, 56, 65, 40, 60] to handle 3D reconstruction tasks while most of these models only approach voxelized 3D objects and do not explore the important role of topology playing in 3D shape reconstruction.

Shape Geometry and Topology. Geometric and topological data analysis [5, 7] has inspired abundant methods to be developed in various directions and applications. Geometric and topological information extracted from object structures provides whole new families of features and descriptors of the data apart from the prior knowledge of 2D images structure. Existing methods [44, 51] have demonstrated effectiveness of geometry and topology components when combined with domain information.

Geometric Constraints within Deep Learning. Many recent studies [8, 23] have shown the potential and capability of geometric and topological properties directly or indirectly helping control and regularize the latent representations by incorporating the topology in meaningful manners. Moreover, geometry-based distances have been recently adopted in point-cloud based 3D object reconstruction [16], object localization [41], and skeleton extraction [64]. While pointing to an important direction to study geometry beyond per-pixel reconstruction losses, these existing methods [16, 41, 64] have been focusing primarily on adding distances such as the Hausdorff that only consider coarse shape matching without an explicit representations to account for fine-grained geometric properties.

Point-cloud and mesh-based approaches. Different from MarrNet which is based on voxel-based output, another line of research aims at producing point cloud (AtlasNet [17]), mesh-based 3D (Pixel2mesh [53] and 3dn [54]), and implicit function based (IM-NET [10], OccNet [34], DISN [63]) reconstructions. While producing encouraging results with a steady improvement, it is not clear how the above methods can handle more topologically challenging data like ABC [28]. For example, 3dn [54] requires an existing 3D template to start with, which is absent in ABC and DISN [63] defines an implicit function for each 3D point being inside, on, or outside the surface and then there exists a potential barrier for DISN to deal with multiple disconnected components of an object. While Skeleton-bridged Deep Learning Approach for Generating Meshes of Complex Topologies [47], adopts different shape representations of point cloud, volume and mesh to recover and refine shapes, their method is not end-to-end trainable.

Relation to MarrNet and ShapeHD. TPWCoder builds on top of MarrNet by adding a 3D autoencoder (AE) with differentiable topological loss. In terms of both visual inspection and numeric metric measures, the improvement of TPWCoder over MarrNet is evident. TPWCoder without the topological loss part already improves over MarrNet. This is understandable since the 3D AE acts as a regularizer, which is lacking in MarrNet. Prior methods exist [37] which use latent variables for feature disentanglement, but they have different formulations and objectives than TPWCoder. TPWCoder with the topological loss is able to correct large errors made by MarrNet [55]. ShapeHD [57] introduces a "naturalness loss" by adding an adversarial term which classifies between real and fake samples. ShapeHD provided certain but limited improvement over MarrNet, which is understandable since the adversarial term is itself not rich enough to account for the explicit topological properties.

3. Method

To introduce a differentiable topological loss, we first explain the topology awareness in general case and thus specifically define two topological properties: genus and connectivity under the ABC dataset [28] scenario. We then introduce our TPWCoder that regularizes the reconstruction shapes from MarrNet and refines local shape details by using topological loss.

3.1. Topological properties

Topology [3], in mathematics, concerns with geometric properties of objects. It captures the key high-level characterization about the 3D object shape and are vital to the understanding and recognition of the object class. Existing methods [16, 41, 64] adding the geometric constraints such as the Hausdorff distances, perform coarse-level correction which are not to maintain the intrinsic topological proper-

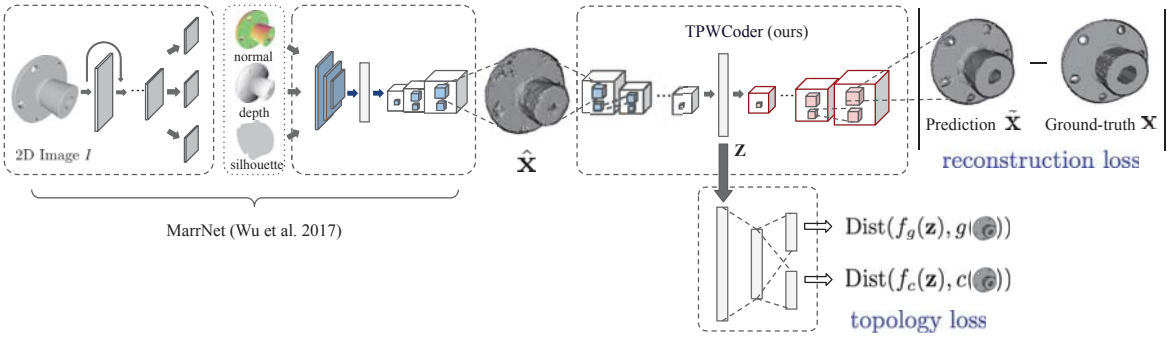

Figure 2: Pipeline of our proposed TPWCoder with decoder. We build our proposed TPWCoder with decoder upon the MarrNet [55].

ties of the objects. We pick two topological properties of 3D geometric objects to study: genus and connectivity.

Genus. The genus of the surface of an object refers to the maximum number of holes, which can be mathematically defined in terms of the Euler characteristic using the Euler's polyhedron formula [45].

$$V - E + F = 2 - 2g \tag{1}$$

where V, E, F are the number of vertices, edges and faces of a 3D object respectively and g is the genus we want to compute. In addition, if an object has more than one component, the genus is defined as the summation of genus of each component.

Connectivity. Connectivity refers to the number of connected parts for the surface of a 3D object. We simply define the connectivity of 3D objects on ABC dataset [28] as the number of connected group of vertices exist in mesh or, in other words, the number of connected components.

Both "genus" and "connectivity" are well defined mathematically [3, 45] and they can be explicitly computed given a mesh-based representation using toolbox like Trimesh [15]. Figure 1 shows some typical examples from the ABC dataset [28] displaying a varying number of holes and connected components. The main challenge is however the computation of genus and connectivity is not differentiable and thus they cannot be directly integrated into the current end-to-end learning framework.

3.2. TPWCoder

To learn computing the topological properties such as genus and connectivity, we propose an autoencoder [22] structure with additional multi-layer neural networks taking the latent variable $\mathbf{z}$ as the input to approximate the topological functions. In [14], it is shown that adding a guidance to the latent variables improves the transparency and disentanglement in the representation learning of the VAE/AE. Figure 4 gives a basic illustration for the architecture. There are three main characteristics for the design of our proposed autoencoder.

1. The latent variable $\mathbf{z}$ is learned as an abstraction for the original 3D shape (in volume) to be used in a neural network to approximate the topological functions that are mathematically specific.

2. The decoder part still tries to reconstruct the original data $\mathbf{X}$ to maintain the regularity of the 3D shape. One could remove the decoder part but our experimental results suggest that having the decoder part being a favorable choice, as seen in Table 1 for the comparison.

3. We call this autoencoder structure, TPWCoder, that can be pretrained using the given 3D shape data and, is differentiable and can be integrated into an end-to-end learning framework as shown in Figure 2.

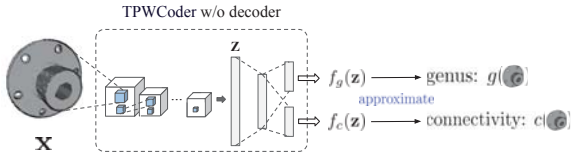

Figure 3: Illustration of our TPWCoder without the decoder.

We specifically design our TPWCoder in two ways: TPWCoder without decoder shown in Figure 3 and TPWCoder with decoder shown in Figure 4.

TPWCoder without decoder. A encoder of five sets of 3D convolutions encodes a $128 \times 128 \times 128$ voxel into a 400-d latent vector $\mathbf{z}$. It is then directly projected to a 100-d vector with a fully-connected layer followed by two sets of 25-way softmax layer for genus g and connectivity c. We formulate the prediction of genus and connectivity as classification tasks in which we limit the value larger than 24 to be 24 for genus and the value larger than 25 to be 25 for connectivity such that $g \in [0, 24]$ and $c \in [1, 25]$.

TPWCoder with decoder. Based on TPWCoder without decoder shown in Figure 3, a decoder of five sets of 3D transposed convolutions is built upon the latent vector $\mathbf{z}$ such that it is upsampled back to $128 \times 128 \times 128$ voxel space for reconstruction purpose.

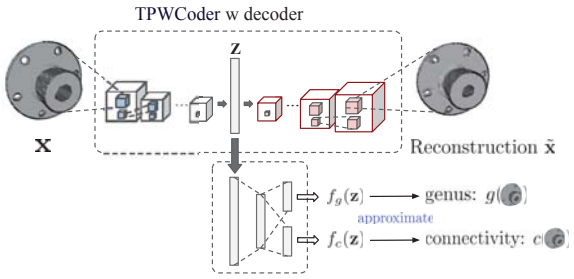

Figure 4: Illustration of our TPWCoder with the decoder.

3.3. End-to-end training

We briefly describe how TPWCoder is built on top of the MarrNet [55], named MarrNet-TPWCoder, for end-to-end training. Given a 2D input image I, our goal is to learn making a prediction for the ground-truth $\mathbf{X}$. Let the reconstruction output of MarrNet be $\hat{\mathbf{X}}$ and the prediction of the final reconstruction output be $\tilde{\mathbf{X}}$. The learning process is trying to minimize the loss:

$$\mathcal{L} = \mathcal{L}_{\text{reconstruction}} + \alpha \mathcal{L}_{\text{topology}} \tag{2}$$

where the first and the second term refer to the reconstruction loss and the topological loss respectively and α is the coefficient balancing the magnitudes of reconstruction loss and topological loss. For reconstruction loss,

$$\mathcal{L}_{\text{reconstruction}} = \text{Dist}(\mathbf{X}, \tilde{\mathbf{X}}) \tag{3}$$

refers to the reconstruction loss trying to minimize the per-voxel difference between the final reconstruction output $\tilde{\mathbf{X}}$ and the ground truth $\mathbf{X}$ in MarrNet-TPWCoder with decoder architecture. $\text{Dist}(\cdot)$ denotes the binary cross-entropy loss in both cases. For topological loss,

$$\mathcal{L}_{\text{topology}} = \text{Dist}(f_g(\mathbf{z}), g(\mathbf{X})) + \text{Dist}(f_c(\mathbf{z}), c(\mathbf{X})),$$

where $g(\mathbf{X})$ and $c(\mathbf{X})$ are respectively the real genus and connectivity numbers for the ground-truth shape $\mathbf{X}$, and $f_g(\mathbf{z})$ and $f_c(\mathbf{z})$ are the predictions by the TPWCoder for the genus and connectivity respectively. $\text{Dist}(\cdot)$ denotes the cross-entropy loss in this case. The loss function $\mathcal{L}$ in Eq. 2 can be minimized via end-to-end training.

Combining TPWCoder with a state-of-the-art 3D shape reconstruction method [55] yields a noticeable performance gain, as shown in Figure 5 and Table 1. The main goal of this paper is learning a faithful TPWCoder to approximate the topological shape properties including but not limited to genus and connectivity. We will try to introduce transparency to the approximation function by studying aspects from [35, 15].

3.4. Training details

We follow the two-step training paradigm described in MarrNet [55] where 2D sketch estimator and 3D shape estimator are trained individually in the first step. We also pre-train our TPWCoder. We then fix the 2D sketch estimator while only fine-tuning 3D shape estimator and TPWCoder using reconstruction loss and topological loss.

2D sketch estimator. We well-train the 2D sketch estimator using ground truth depth, normal and silhouette with $L2$ loss, a batch size of 64 and Adam optimizer with a learning rate of 10^{-3} for 200 epochs. Well-trained 2D sketch estimator is then fixed during fine-tuning.

3D shape estimator. We pre-train the 3D shape estimator using ground truth depth, normal, silhouette and ground truth voxel as guidance with binary cross-entropy loss, a batch size of 16 and Adam optimizer with a learning rate of 10^{-3} for 40 epochs.

TPWCoder. We pre-train our TPWCoder using ground truth voxel, genus and connectivity along with 3D shape reconstructions from MarrNet [55] and their corresponding genus and connectivity as guidance with binary cross-entropy loss and topological loss stated in Eq. (2). A batch size of 32 and Adam optimizer with a learning rate of 10^{-4} are used for training 100 epochs.

At fine-tuning stage, we fix the 2D sketch estimator and fine-tune the 3D shape estimator and TPWCoder with both reconstruction loss and topological loss shown in Eq. (2). We fine-tune the network using a batch size of 16 and Adam optimizer with a learning rate of 10^{-4} for 80 epochs.

4. Experiments

In this section, we evaluate variants of MarrNet-TPWCoder on ABC dataset [6] in our ablation study. We then compare the reconstruction performance of MarrNet-TPWCoder with MarrNet [55] and ShapeHD [57] quantitatively and qualitatively on ABC dataset [28] and ShapeNet dataset [6]. Moreover, we compare the real image 3D reconstruction of MarrNet-TPWCoder, MarrNet [55] and ShapeHD [57] on Pascal 3D+ dataset [59] qualitatively. We explore the effectiveness of our TPWCoder in handling the topology on ABC dataset [28] at the end.

Datasets. In order to evaluate our TPWCoder in managing topology and reconstruction, we use the ABC dataset [28] which has rich topological features. For fair comparison, we make use of ShapeNet datasets [6], specifically chair, car and plane classes.

For the ABC dataset [28] where each object does not have class property, we make use of 22826 objects and use Blender [12] to render corresponding RGB, normal, depth

and silhouette of each object in viewer-centered. We also use a mesh-based library, Trimesh [15], to generate the ground truth voxel and corresponding genus and connectivity of each object. We then randomly split 22826 objects into 20544 objects for training and the rest for testing.

For the ShapeNet dataset [57], we adopt the exact same rendering strategy as MarrNet [55] and ShapeHD [57].

Metrics. To fully evaluate the 3D shape reconstruction quantitatively, we use three standard metrics: Intersection of Union (IoU), Chamfer Distance (CD) and Earth Mover's Distance (EMD), for evaluating voxel-based reconstruction and two self-defined metrics: Genus Number Error (GNE) and Connectivity Number Error (CNE), for reconstructed shape topology analysis.

For IoU, CD and EMD, we follow the same evaluation details as the implementation in Pix3D [46]. We define the Genus Number Error (GNE) and Connectivity Number Error (CNE) in terms of the genus and connectivity of reconstructed object and its corresponding ground truth object:

$$E = \frac{1}{n}\sum_{i=1}^{n} |g(\tilde{\mathbf{X}}_i) - g(\mathbf{X}_i)| \tag{4}$$

where E denotes the GNE or CNE, $g(\tilde{\mathbf{X}}_i)$ and $g(\mathbf{X}_i)$ are genus/connectivity of reconstructed and ground truth object respectively and n is the number of objects.

4.1. Results on ABC

We first quantitatively compare the reconstruction results of MarrNet-TPWCoder without decoder and with decoder. For each case, we choose to fix the TPWCoder and fine-tune the TPWCoder with topological loss therefore forms four comparison experiments. We set the coefficient α shown in Eq. (2) as 10^{-3} for balancing the magnitudes of reconstruction loss and topological loss. Table 1 provides quantitative comparison. Notice that for MarrNet-TPWCoder without decoder, fine-tuning makes noticeable improvement than without fine-tuning in CD, EMD and CNE, same trend happens in MarrNet-TPWCoder with decoder as well. Moreover, MarrNet-TPWCoder with decoder outperforms without decoder significantly in IoU, CD, EMD and CNE and improves GNE. We therefore make use of fine-tuned MarrNet-TPWCoder with decoder to compare with MarrNet [55] and ShapeHD [57].

We then compare the MarrNet-TPWCoder with MarrNet [55] and ShapeHD [57]. Additionally, we trained MarrNet-TPWCoder without using topological loss for full comparison. The quantitative results comparison is provided in Table 1. Note that MarrNet-TPWCoder outperforms MarrNet [55] and ShapeHD [57] by a noticeable margin.

We also present qualitative results comparison in Figure 5. It is noticeable that both MarrNet-TPWCoder without topology and MarrNet-TPWCoder with topology produce much distinct structures, smoother surfaces and cleaner edges than MarrNet [55] and ShapeHD [57]. In particular, the reconstructed object surface generated by MarrNet-TPWCoder with topology is visibly smoother than MarrNet-TPWCoder without topology and more unnecessary particles are eliminated, producing cleaner reconstruction results.

4.2. TPWCoder Topology-Awareness

We show the capability of our topology-aware TPWCoder in predicting genus and connectivity on ABC dataset [28]. Figure 6 demonstrates the topology prediction error along with the samples distribution of genus and connectivity on the testing set of ABC dataset [28].

Note that either the genus prediction error or connectivity prediction error of each class is computed by summing up the distance between predicted class and ground truth class of each object in that class. That is, the further the prediction is away from truth value, the larger the error would be. For the genus, we then compare the prediction error with the error from all 0 predictions, similarly for the connectivity, we compare the prediction error with the error from all 1 predictions in that genus $g = 0$ and connectivity $c = 1$ dominate the genus and connectivity respectively. As shown in Figure 6, TPWCoder attempts to capture the topological information through encoder and latent representation and make prediction instead of randomly guessing or predicting constant, especially when an object has genus or connectivity less than 5.

Note also that for the connectivity, there are some prediction faultages in few specific classes therefore the prediction error curve overlaps with all 1 prediction error. This is because some classes only have very few samples or even no sample in the testing set.

4.3. Results on ShapeNet

We now compare the reconstruction results of MarrNet-TPWCoder with MarrNet [55] and ShapeHD [57] on ShapeNet dataset [6]. The coefficient α shown in Eq. (2) is set to 10^{-3} for balancing the magnitudes of reconstruction loss and topological loss.

The quantitative results are presented in Table 2. For chair class, ShapeHD [57] yields better results than ours. MarrNet-TPWCoder yields better results in both car and plane class. Note, although, that the ShapeNet dataset [6] does not have much topological information therefore TPWCoder possibly does not well refine local details of shapes, MarrNet-TPWCoder still produces better quantitative results. We provide qualitative results in Figure 7.

4.4. Results on Pascal 3D+

As suggested in [59], Pascal 3D+ should not be used for benchmark 3D object reconstruction training since all ob-

Table 1: Quantitative comparison on the ABC dataset [28]. IoU: Intersection over Union; CD: Chamfer Distance; EMD: Earth Mover's Distance; GNE: Genus Number Error; CNE: Connectivity Number Error. ↑ means the higher the better; ↓ means the lower the better.

METHODS			IoU ↑	CD ↓	EMD ↓	GNE ↓	CNE ↓
MarrNet [55]	-	-	0.6132	0.0916	0.0957	16.0158	13.5215
ShapeHD [57]	-	-	0.6168	0.0858	0.0869	16.2551	13.0964
MarrNet-TPWCoder w/o L_{topology}	w Decoder	w Fine-tune	0.6252	**0.0770**	0.0791	**14.2940**	**11.6725**
MarrNet-TPWCoder w L_{topology}	w/o Decoder	w/o Fine-tune	0.5801	0.1508	0.1455	15.2117	20.3732
		w Fine-tune	0.6023	0.1139	0.1160	15.6276	15.2421
	w Decoder	w/o Fine-tune	0.6152	0.0930	0.0968	14.4729	13.9437
		w Fine-tune	**0.6313**	0.0792	**0.0784**	14.5915	12.8067

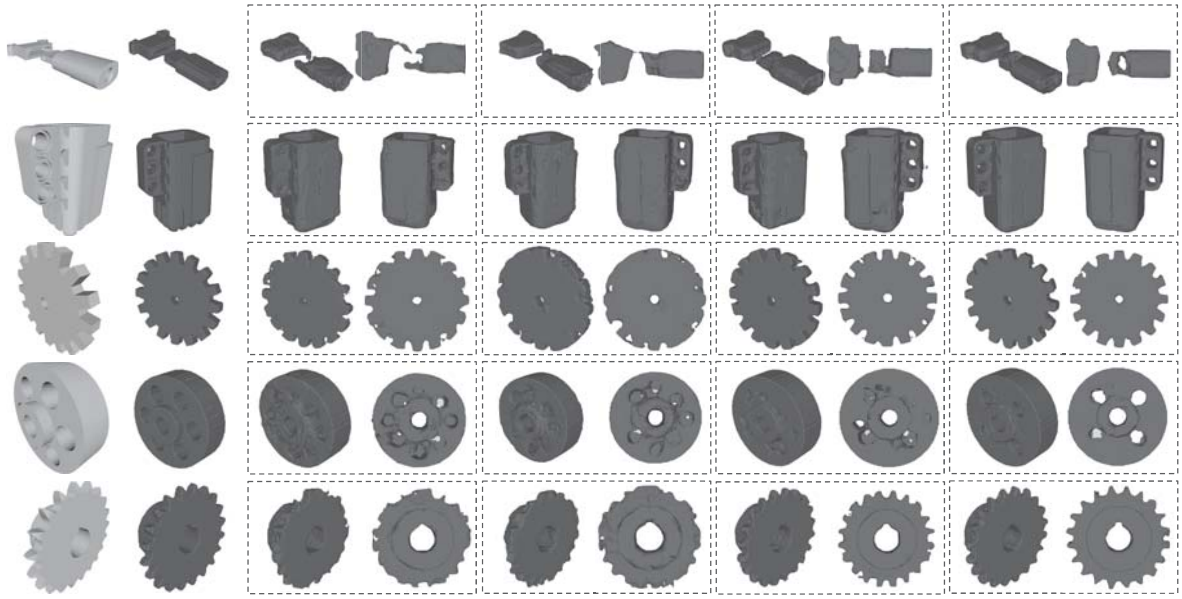

Figure 5: Reconstruction results for 5 examples from the ABC dataset [28] with different approaches: MarrNet, ShapeHD and our method using TPWCoder without and with topological loss.

Table 2: Quantitative comparison on the ShapeNet dataset [6]. IoU: Intersection over Union; CD: Chamfer Distance; EMD: Earth Mover's Distance. ↑ means the higher the better; ↓ means the lower the better.

METHODS	IoU ↑			CD ↓			EMD ↓		
	chair	car	plane	chair	car	plane	chair	car	plane
MarrNet [55]	**0.4482**	0.6542	0.4496	0.1838	0.0766	0.0810	0.1761	0.0803	0.0998
ShapeHD [57]	0.4056	0.6584	0.4676	**0.1779**	0.0776	0.0778	**0.1624**	0.0791	0.0989
MarrNet-TPWCoder w/o L_{topology} (ours)	0.4327	0.6829	**0.5502**	0.2260	0.0716	**0.0763**	0.2072	0.0758	**0.0931**
MarrNet-TPWCoder w L_{topology} (ours)	0.4360	**0.6953**	0.5488	0.2286	**0.0681**	0.0782	0.2086	**0.0728**	0.0968

jects in both training set and testing set share the same 10 CAD model thus reconstruction would be biased. In this case, we directly test MarrNet-TPWCoder without topology, MarrNet-TPWCoder with topology, MarrNet [55] and ShapeHD [57], which are well-trained on ShapeNet dataset [6], on Pascal 3D+ dataset [59] and qualitatively compare the reconstruction results.

Real images reconstruction results comparison are provided in Figure 8. MarrNet-TPWCoder produces much more intact shape (*i.e.* the chair leg, the contour of car and

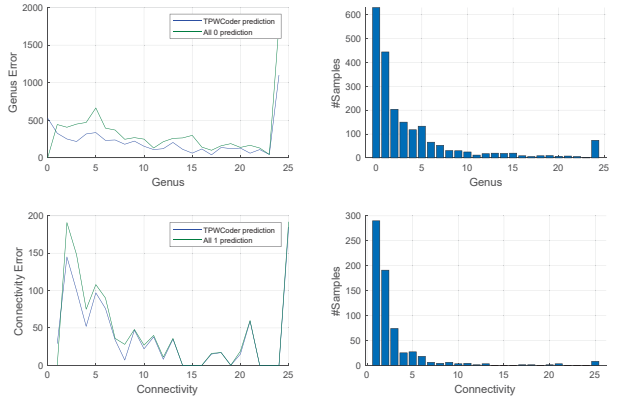

Figure 6: Performance of TPWCoder in predicting genus and connectivity. **Left**: Genus prediction error curve and connectivity prediction error curve. **Right**: Test sample distribution of genus and connectivity.

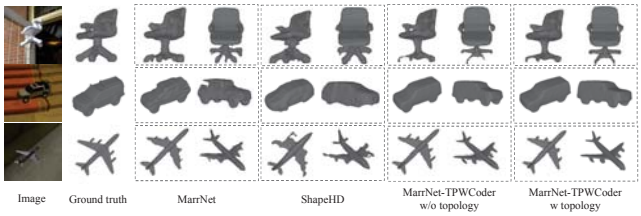

Figure 7: Reconstruction examples from the ShapeNet dataset [28] with different approaches: MarrNet, ShapeHD and our method using TPWCoder without and with topological loss.

aerofoil and engines). MarrNet-TPWCoder with topological loss further improves the details of reconstructed shape (*i.e*. surfaces and edges) compared to MarrNet-TPWCoder without topological loss.

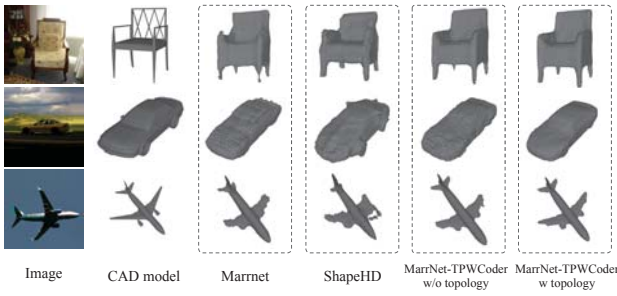

Figure 8: Reconstruction results for 3 examples from the Pascal 3D+ dataset [59] with different approaches: MarrNet, ShapeHD and our method using TPWCoder without and with topological loss.

5. Conclusion

In this paper, we make an attempt to address topology-awareness for 3D shape reconstruction by proposing topology-aware shape autoencoder, TPWCoder, that learns to approximate 3D topological functions as well as to reconstruct 3D shapes. We have shown how to combine our TPWCoder with other architectures, *i.e*. MarrNet [55], therefore could be built upon other architectures for different tasks with customized modification. Our ablation study of variants of TPWCoder on the ABC dataset [28] has shown that decoder architecture and fine-tuning stage are both needed to produce better reconstruction results in that TPWCoder is capable of regularizing global shape reconstruction and topological loss helps in refining the local details. Our experiments on ABC dataset [28] and ShapeNet dataset [6] have demonstrated that TPWCoder makes its attempts to address topology-awareness and is able to refine local details of reconstruction as well.

6. Acknowledgements

Our work is supported by NSF IIS1618477 and IIS-1717431. We thank Jiajun Wu, Kwonjoon Lee, Weijian Xu, Justin Lazarow, Sainan Liu, and Hao Su for valuable discussions.

References

[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011. 1

[2] Adelbert Ames Jr. Visual perception and the rotating trapezoidal window. *Psychological Monographs: General and Applied*, 65(7):i, 1951. 1

[3] Glen E Bredon. *Topology and geometry*, volume 139. Springer Science & Business Media, 2013. 2, 3

[4] Myron Z Brown, Darius Burschka, and Gregory D Hager. Advances in computational stereo. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (8):993–1008, 2003. 1

[5] Gunnar Carlsson. Topology and data. Technical report, 2008. 2

[6] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012, Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 2, 4, 5, 6, 7

[7] Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *arXiv preprint arXiv:1710.04019*, 2017. 2

[8] Chao Chen, Xiuyan Ni, Qinxun Bai, and Yusu Wang. A topological regularizer for classifiers via persistent homology. *arXiv preprint arXiv:1806.10714*, 2018. 2

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image

segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018. 1

[10] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, 2019. 2

[11] Ismael Colomina and Pere Molina. Unmanned aerial systems for photogrammetry and remote sensing: A review. *ISPRS Journal of photogrammetry and remote sensing*, 92:79–97, 2014. 1

[12] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 4

[13] Daniel Cremers, Mikael Rousson, and Rachid Deriche. A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape. *International journal of computer vision*, 72(2):195–215, 2007. 1

[14] Zheng Ding, Yifan Xu, Weijian Xu, Yang Yang, Max Welling, and Zhuowen Tu. Guided variational autoencoder for disentanglement learning. In *CVPR*, 2020. 3

[15] Dawson Haggerty *et al*. Trimesh [computer software]. (2019). retrieved from https://github.com/mikedh/trimesh. 3, 4, 5

[16] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *CVPR*, 2017. 2

[17] Thibault Groueix, Matthew Fisher, Vladimir Kim, Bryan Russell, and Mathieu Aubry. Atlasnet: A papier-mâché approach to learning 3d surface generation. In *CVPR 2018*, 2018. 2

[18] Feng Han and Song-Chun Zhu. Bayesian reconstruction of 3d shapes and scenes from a single image. In *First IEEE International Workshop on Higher-Level Knowledge in 3D Modeling and Motion Analysis, 2003. HLK 2003.*, 2003. 1

[19] Adam P Harrison, Ziyue Xu, Kevin George, Le Lu, Ronald M Summers, and Daniel J Mollura. Progressive and multi-path holistically nested neural networks for pathological lung segmentation from ct images. In *MICCAI*, pages 621–629, 2017. 1

[20] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1

[21] Harold C Hill and Alan Johnston. The hollow-face illusion: Object specific knowledge, general assumptions or properties of the stimulus. 2007. 1

[22] Geoffrey E Hinton and Richard S Zemel. Autoencoders, minimum description length and helmholtz free energy. In *Advances in neural information processing systems*, pages 3–10, 1994. 3

[23] Christoph Hofer, Roland Kwitt, Marc Niethammer, and Mandar Dixit. Connectivity-optimized representation learning via persistent homology. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 2751–2760, 2019. 2

[24] Qixing Huang, Hai Wang, and Vladlen Koltun. Single-view reconstruction via joint analysis of image and shape collections. *ACM Transactions on Graphics*, 34, 2015. 2

[25] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1

[26] Abhishek Kar, Shubham Tulsiani, João Carreira, and Jitendra Malik. Category-specific object reconstruction from a single image. *CVPR*, pages 1966–1974, 2015. 2

[27] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, 2016. 1

[28] Sebastian Koch, Albert Matveev, Zhongshi Jiang, Francis Williams, Alexey Artemov, Evgeny Burnaev, Marc Alexa, Denis Zorin, and Daniele Panozzo. Abc: A big cad model dataset for geometric deep learning. In *CVPR*, 2019. 1, 2, 3, 4, 5, 6, 7

[29] Kiriakos N Kutulakos and Steven M Seitz. A theory of shape by space carving. *International journal of computer vision*, 38(3):199–218, 2000. 1

[30] Wilfried Linder. *Digital photogrammetry*. Springer, 2009. 1

[31] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2024–2039, 2015. 1

[32] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1

[33] Yi Ma, Stefano Soatto, Jana Kosecka, and S Shankar Sastry. *An invitation to 3-d vision: from images to geometric models*, volume 26. Springer Science & Business Media, 2012. 1

[34] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 2

[35] Bojan Mohar and Carsten Thomassen. *Graphs on surfaces*, volume 2. Johns Hopkins University Press Baltimore, 2001. 4

[36] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew W Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *ISMAR*, volume 11, pages 127–136, 2011. 1

[37] Xi Peng, Xiang Yu, Kihyuk Sohn, Dimitris N Metaxas, and Manmohan Chandraker. Reconstruction-based disentanglement for pose-invariant face recognition. In *ICCV*, 2017. 2

[38] Marc Pollefeys, Reinhard Koch, and Luc Van Gool. Self-calibration and metric reconstruction inspite of varying and unknown intrinsic camera parameters. *International Journal of Computer Vision*, 32(1):7–25, 1999. 1

[39] Marc Pollefeys, David Nistér, J-M Frahm, Amir Akbarzadeh, Philippos Mordohai, Brian Clipp, Chris Engels, David Gallup, S-J Kim, Paul Merrell, et al. Detailed real-time urban 3d reconstruction from video. *International Journal of Computer Vision*, 78(2-3):143–167, 2008. 1

[40] Danilo Jimenez Rezende, S. M. Ali Eslami, Shakir Mohamed, Peter Battaglia, Max Jaderberg, and Nicolas Heess. Unsupervised learning of 3d structure from images, 2016. 2

[41] Javier Ribera, David Güera, Yuhao Chen, and Edward Delp. Weighted hausdorff distance: A loss function for object localization. *arXiv preprint arXiv:1806.07564*, 2018. 2

[42] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *CVPR*, 2017. 2

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 1

[44] Gurjeet Singh, Facundo Mémoli, and Gunnar E. Carlsson. Topological methods for the analysis of high dimensional data sets and 3d object recognition. In *SPBG*, 2007. 2

[45] Edwin H Spanier. *Algebraic topology*. Springer Science & Business Media, 1989. 3

[46] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *CVPR*, 2018. 5

[47] Jiapeng Tang, Xiaoguang Han, Junyi Pan, Kui Jia, and Xin Tong. A skeleton-bridged deep learning approach for generating meshes of complex topologies from single rgb images, 2019. 1, 2

[48] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *ICCV*, 2017. 2

[49] Shubham Tulsiani, Saurabh Gupta, David F Fouhey, Alexei A Efros, and Jitendra Malik. Factoring shape, pose, and layout from the 2d image of a 3d scene. In *CVPR*, 2018. 1

[50] Shubham Tulsiani, Tinghui Zhou, Alexei A. Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *CVPR*, 2017. 2

[51] Katharine Turner, Sayan Mukherjee, and Doug M Boyer. Persistent homology transform for modeling shapes and surfaces. *arXiv preprint arXiv:1310.1030*, 2013. 2

[52] Shimon Ullman. The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 203(1153):405–426, 1979. 1

[53] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*, 2018. 2

[54] Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. 3dn: 3d deformation network. In *CVPR*, 2019. 2

[55] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum. Marrnet: 3d shape reconstruction via 2.5 d sketches. In *Advances in neural information processing systems*, 2017. 1, 2, 3, 4, 5, 6, 7

[56] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T Freeman, and Joshua B Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in Neural Information Processing Systems*, 2016. 2

[57] Jiajun Wu, Chengkai Zhang, Xiuming Zhang, Zhoutong Zhang, William T Freeman, and Joshua B Tenenbaum. Learning shape priors for single-view 3d completion and reconstruction. In *ECCV*, 2018. 1, 2, 4, 5, 6

[58] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, 2015. 1

[59] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *WACV*, March 2014. 4, 5, 6, 7

[60] Jianwen Xie, Zilong Zheng, Ruiqi Gao, Wenguan Wang, Song-Chun Zhu, and Ying Nian Wu. Learning descriptor networks for 3d shape synthesis and analysis. In *CVPR*, 2018. 2

[61] Saining Xie, Xun Huang, and Zhuowen Tu. Top-down learning for structured labeling with convolutional pseudoprior. In *ECCV*, 2016. 1

[62] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, 2015. 1

[63] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In *Advances in Neural Information Processing Systems 32*, pages 492–502. Curran Associates, Inc., 2019. 1, 2

[64] Weijian Xu, Gaurav Parmar, and Zhuowen Tu. Learning geometry-aware skeleton detection. In *BMVC*, 2019. 2

[65] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *Advances in Neural Information Processing Systems*. 2016. 2

[66] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017. 1

[67] Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Josh Tenenbaum, Bill Freeman, and Jiajun Wu. Learning to reconstruct shapes from unseen classes. In *Advances in Neural Information Processing Systems 31*. 2018. 1

[68] Song-Chun Zhu, David Mumford, et al. A stochastic grammar of images. *Foundations and Trends® in Computer Graphics and Vision*, 2(4):259–362, 2007. 1

[69] Michael Zollhöfer, Patrick Stotko, Andreas Görlitz, Christian Theobalt, Matthias Nießner, Reinhard Klein, and Andreas Kolb. State of the art on 3d reconstruction with rgb-d cameras. In *Computer graphics forum*, volume 37, pages 625–652. Wiley Online Library, 2018. 1