Enhanced 3D Shape Reconstruction With Knowledge Graph of Category Concept

GUOFEI SUN, Zhejiang University, China
YONGKANG WONG and MOHAN S. KANKANHALLI, National University of Singapore
XIANGDONG LI and WEIDONG GENG, Zhejiang University, China

Reconstructing three-dimensional (3D) objects from images has attracted increasing attention due to its wide applications in computer vision and robotic tasks. Despite the promising progress of recent deep learning–based approaches, which directly reconstruct the full 3D shape without considering the conceptual knowledge of the object categories, existing models have limited usage and usually create unrealistic shapes. 3D objects have multiple forms of representation, such as 3D volume, conceptual knowledge, and so on. In this work, we show that the conceptual knowledge for a category of objects, which represents objects as prototype volumes and is structured by graph, can enhance the 3D reconstruction pipeline. We propose a novel multimodal framework that explicitly combines graph-based conceptual knowledge with deep neural networks for 3D shape reconstruction from a single RGB image. Our approach represents conceptual knowledge of a specific category as a structure-based knowledge graph. Specifically, conceptual knowledge acts as visual priors and spatial relationships to assist the 3D reconstruction framework to create realistic 3D shapes with enhanced details. Our 3D reconstruction framework takes an image as input. It first predicts the conceptual knowledge of the object in the image, then generates a 3D object based on the input image and the predicted conceptual knowledge. The generated 3D object satisfies the following requirements: (1) it is consistent with the predicted graph in concept, and (2) consistent with the input image in geometry. Extensive experiments on public datasets (i.e., ShapeNet, Pix3D, and Pascal3D+) with 13 object categories show that (1) our method outperforms the state-of-the-art methods, (2) our prototype volume-based conceptual knowledge representation is more effective, and (3) our pipeline-agnostic approach can enhance the reconstruction quality of various 3D shape reconstruction pipelines.

CCS Concepts: • **Computing methodologies** → **Volumetric models**; *Neural networks;*

Additional Key Words and Phrases: Deep learning, 3D reconstruction, conceptual knowledge

This work is supported by the National Key Research and Development Program of China (No. 2017YFB1303201 and No. 2017YFB1002802), and the National Natural Science Foundation of China (No. 61972346).
Authors' addresses: G. Sun and W. Geng (corresponding author), State Key Laboratory of CAD&CG, College of Computer Science and Technology, Zhejiang University, Zheda Road No. 38, Hangzhou, Zhejiang, 310027, China; emails: {guofeisun, gengwd}@zju.edu.cn; Y. Wong, School of Computing, National University of Singapore, 3 Research Link, i4.0 Building, Singapore 117602; email: yongkang.wong@nus.edu.sg; M. S. Kankanhalli, School of Computing, National University of Singapore, 11 Computing Drive, AS6 Building, 117416, Singapore; email: mohan@comp.nus.edu.sg; X. Li (corresponding author), College of Computer Science and Technology, Zhejiang University, Zheda Road No. 38, Hangzhou, Zhejiang, 310027, China; email: axli@zju.edu.cn.

1551-6857/2022/02-ART71 $15.00
https://doi.org/10.1145/3491224

ACM Reference format:
Guofei Sun, Yongkang Wong, Mohan S. Kankanhalli, Xiangdong Li, and Weidong Geng. 2022. Enhanced 3D Shape Reconstruction With Knowledge Graph of Category Concept. *ACM Trans. Multimedia Comput. Commun. Appl.* 18, 3, Article 71 (February 2022), 20 pages.
https://doi.org/10.1145/3491224

1 INTRODUCTION

Reconstructing an accurate three-dimensional (3D) object from a single-view image is a fundamental computer vision task that can further improve the performance of high-level vision tasks, such as virtual reality/augmented reality (VR/AR), object localization, shape retrieval and deformation, and so on. However, reconstructing 3D shapes from 2D images is very difficult since it is an under-constrained problem. To generate 3D shapes from single-view images, existing methods use silhouettes [8] or shading [43] to infer 3D shapes. Due to the strong assumptions in these methods, their deployment in real-world applications remains challenging.

With the recent advance in deep learning research [29], 3D shape reconstruction from single-view images has made great progress with deep neural networks. Currently, most approaches reconstruct 3D shapes based on two frameworks. The first is an encoder-decoder framework, in which 2D convolution layers (i.e., encoder) are used to extract features from the input image and 3D deconvolution layers (i.e., decoder) are used to reconstruct the corresponding 3D shape. The second is an encoder-decoder-refiner framework, in which the 3D convolution and deconvolution layers (i.e., refiner) are used to refine the reconstructed 3D shape from the encoder-decoder part. To further improve reconstruction performance, some approaches added other prior information in the reconstruction pipeline, such as view prior [26] and shape prior [63, 72]. With the additional prior information during training or inference stage, the reconstructed 3D shapes are more accurate.

Despite the promising results from the previous methods [26, 63], there remain some limitations to deal with. For example, some of the generated shapes are unrealistic and the fine-grained details of small parts are ignored. This is mainly because the reconstruction pipeline accepts only 2D images with low-level priors as inputs. In fact, 3D shapes are composed of many conceptual parts [41]. For example, a chair is composed of base, seat, arms, and back rest. As shown in this article, these conceptual parts help improve reconstruction quality. To the best of our knowledge, this is the first time this kind of knowledge is employed in the single-view 3D reconstruction task.

When humans perceive an image, they can predict the 3D shape of the object in the image and construct the conceptual knowledge about this object (e.g., this object is a chair, it has a base, a seat, two arms, and a back rest) based on experience. Humans can also effortlessly come up with various novel shapes from a single image by removing or replacing some conceptual parts. Motivated by these observations, we propose a novel framework that explicitly combines graph-based conceptual knowledge with deep neural networks for 3D shape reconstruction from a single RGB image. As shown in Figure 1, our framework takes an RGB image as input, predicts the conceptual knowledge from the image, and generates a 3D object based on the input image and the predicted conceptual knowledge. The generated 3D object satisfies the following requirements: (1) it is consistent with the predicted conceptual knowledge graph, and (2) consistent with the geometric of the input image. Our framework is composed of six modules: *concept classifier (image), image encoder*, *conceptual knowledge encoder*, *volume decoder*, *volume refiner*, and *concept classifier (volume).* The concept classifier (image) predicts conceptual knowledge (i.e., object category and conceptual part labels) from the input image. The image encoder and conceptual knowledge encoder are used to extract features from the given input image and predicted conceptual knowledge, respectively.

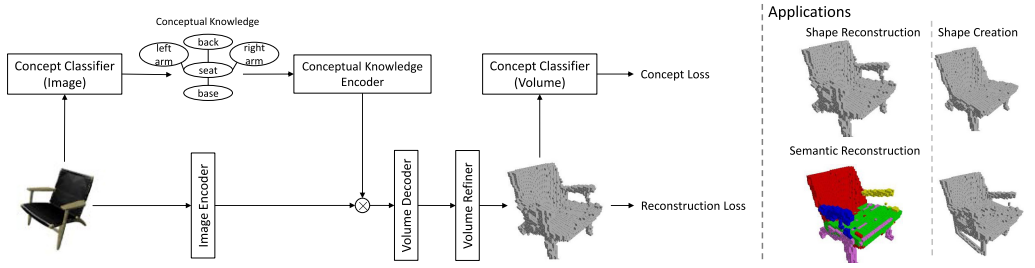

Fig. 1. An overview of the proposed approach (left) and applications enabled by this approach (right).

The extracted features are then fused and fed to the volume decoder and volume refiner to reconstruct a concept-aware 3D shape. The concept classifier (volume) extracts conceptual part labels from the reconstructed 3D shape, which is used in concept loss for concept consistency between the input conceptual knowledge and the reconstructed 3D shape. In addition to concept loss, reconstruction loss is employed to impose geometric consistency.

By explicitly incorporating graph-based conceptual knowledge, our framework also enables several novel applications. First, our framework can reconstruct 3D shapes from a single RGB image without any additional information, which is the standard single-view 3D reconstruction pipeline. Second, by manually changing the predicted conceptual knowledge, our framework can be used in shape creation tasks. Our framework can create various reasonable 3D shapes from a single image with different conceptual knowledge, which mimics how humans can conceptualize various shapes from the given images. When the conceptual knowledge is decomposed into different sub-types, our proposed method can create novel 3D objects from multiple images and the corresponding sub-types. Finally, our framework can also perform the semantic reconstruction task via concept-aware shape reconstruction and shape subtraction.

The contributions of this work are summarized as follows:

(1) We propose a novel multimodal framework that explicitly combines conceptual knowledge of object category with deep neural networks for 3D shape reconstruction from a single RGB image. In our framework, we propose a novel conceptual knowledge encoder and two concept classifiers that can be directly combined with current 3D reconstruction frameworks. Conceptual knowledge not only helps to reconstruct more accurate 3D shapes from images but can also guide the creation of novel 3D shapes.
(2) We design a concept loss that penalizes the structural dissimilarity between reconstructed shape and the conceptual knowledge for category-specific shape reconstruction. The concept loss provides a global understanding of the reconstructed 3D shapes and improves reconstruction quality by enforcing the network to reconstruct 3D shapes that are similar to the predicted conceptual knowledge.
(3) Experimental results on 13 object categories from ShapeNet, Pascal3D+, and Pix3D demonstrate that our framework outperforms the state-of-the-art methods in 3D reconstruction tasks. As a pipeline-agnostic framework, our approach also enhances the reconstruction quality of other 3D reconstruction pipelines. With the assistance of conceptual knowledge, our framework can enable several applications, such as shape creation and semantic shape reconstruction.

The rest of the article is organized as follows. We review the related work in Section 2. In Section 3, we elaborate on the proposed 3D reconstruction framework that explicitly combines graph-based conceptual knowledge with deep neural networks. Section 4 delineates the experiment

details, results, ablation studies, and additional novel applications enabled by the proposed framework; it also discusses the limitations of this framework. We present our conclusions in Section 5.

2 RELATED WORK

In this section, we review 3D shape reconstruction methods related to our work, which can be categorized into three parts: (1) traditional 3D shape reconstruction, (2) deep learning–based 3D shape reconstruction, and (3) deep learning–based 3D shape generation.

2.1 Traditional 3D Shape Reconstruction

Reconstructing 3D shapes from a single-view image is a challenging task as it is an ill-posed problem. At the early research stage, most approaches infer 3D information of the object from multiple images. SfM [59] and SLAM [12] are widely used in this area. These approaches first extract pixel-level features and match the pixels in RGB image sequences. Then, the relative poses between images are computed by minimizing the reprojection error [4]. Finally, the 3D information is acquired using multi-view geometry [19]. KinectFusion [21, 36] and its follow-ups [75–77] are proposed to reconstruct 3D rigid objects from dense depth image sequences. Some approaches also use depth image sequences to reconstruct human faces [38] and human bodies [2]. These approaches have a few limitations. The pixel feature extraction is compute intensive, and the matching and camera tracking can easily fail when the viewpoints of images are sparse. In addition, to reconstruct the whole object, these approaches need to acquire image collections of all surfaces of the object, which is not feasible and is expensive in some scenarios.

To reconstruct a 3D shape from a single image, previous approaches attempt to address this ill-posed problem using additional clues, such as shading [43], silhouette [8], and texture [60]. However, these approaches require strong assumptions and are difficult to apply in real-world scenarios. Kar et al. [25] focused on category-specific reconstruction, which reconstructs the 3D object of a specific category through deforming the mean shape. It can reconstruct reasonable shapes but is limited to specific categories.

2.2 Deep Learning–Based 3D Shape Reconstruction

With the recent advance of deep learning [29] and the release of large 3D-shape datasets (e.g., ShapeNet [5]), many deep learning–based approaches have achieved promising performance in 3D-shape reconstruction tasks. 3D-R2N2 [10] and LSM [24] reconstruct 3D volume from single and multiple RGB images. A recurrent neural network (RNN)–based fusion module is adopted in these works to fuse features from images. Considering the time-consuming process of RNN, some follow-up approaches fuse image features using max pooling [56] and context-aware fusion [67].

These methods can reconstruct 3D shapes from single and multiple images, while many other methods focus on single-view 3D reconstruction. Wu et al. [64] proposed a convolutional deep belief network to reconstruct 3D volume from a single-depth image. MarrNet [61] first predicted 2.5D sketches (i.e., normal, depth, and silhouette) from an RGB image, then reconstructed a 3D shape using the predicted sketches. To preserve naturalness of the generated shape, Wu et al. [63] developed naturalness loss using an adversarial model based on MarrNet. 3D-RecGAN++ [70] adopted a conditional adversary framework to reconstruct a 3D shape from a single-depth image. It proposed a mean feature discriminator to stabilize the adversary training process. Tulsiani et al. [51] presented a framework that uses multi-view consistency as a supervision signal for single-view reconstruction. Liao et al. [31] disentangled shape feature and pose feature from an image, and reconstructed a 3D shape via 2D-3D self-consistency constraints. Wang et al. [53] combined visual hull with 3D shape reconstruction to refine coarse volumes with predicted visual hull. Wang et al. [54] predicted the depth map from an input RGB image, and projected it as a partial volume. The final 3D

shape was reconstructed via volume refinement. These approaches attempted to improve reconstruction quality with priors and constraints, which are low-level information and lead to limited improvements. Our approach uses high-level information (e.g., conceptual knowledge), which provides a global understanding of the 3D shapes and achieves significant improvement. High-level information also enables our approach for various applications, such as semantic reconstruction, shape creation, and more. To reconstruct high-resolution 3D volumes and reduce memory costs, Richter and Roth [44] proposed shape layer and voxel tube structure. Other approaches, such as HSP [18] and OGN [50], adopted octree [33] to represent the 3D shape and modify the traditional convolution layer to adapt to the octree representation.

According to [17], there are mainly two categories of 3D shape reconstruction tasks: volumetric methods and surface-based methods. Besides the volumetric methods mentioned earlier, the surface-based methods reconstruct 3D shapes using representations such as point clouds [3, 13, 28, 32, 55] and 3D mesh [14, 15, 23, 26, 46, 49, 73, 74]. The final 3D shapes are obtained by deforming the points of a template point cloud [9, 37] or a template 3D mesh [15, 39, 40, 57]. In this work, we focus on the volumetric shape reconstruction task and leave the surface-based shape reconstruction task as future work.

Considering the fact that collecting 3D shape ground- truth is difficult and expensive, some methods attempt to reconstruct 3D shapes without 3D supervision. Some works applied projection loss that used 2D silhouettes as supervision signals to train the network [20, 35, 69, 71]. The reconstructed 3D shape is projected to the 2D image plane via a differentiable projection module such that the loss can be computed by calculating the differences between the projected silhouette and the ground-truth silhouette. Alternately, Xiang et al. [65] used multi-view normal maps as a supervision signal, which integrates pose estimation and mesh reconstruction in the same optimization procedure.

2.3 Deep Learning–Based 3D Shape Generation and Completion

3D shape generation and completion aims at generating reasonable 3D shapes from random feature vectors and partial 3D shapes. Wu et al. [62] proposed 3D-GAN and 3D-VAE-GAN models to generate category-specific 3D shapes from a 200-dimensional vector. Chen and Zhang [7] proposed the IM-NET model to improve the visual quality of the generated shapes. They designed a novel implicit decoder that accepts feature vector and point coordinates as input to generate high-resolution 3D shapes. Han et al. [16] used deep neural networks to infer the global structure of the partial 3D shape and complete the shape through local geometry refinement. Dai et al. [11] first predicted coarse shapes from partial input. Then, the details of the coarse shape were filled using geometric prior from the database. Wang et al. [58] treated the shape completion task as 3D in-painting. They combined 3D encoder-decoder GAN and long short-term memory (LSTM) to complete high-resolution 3D shapes from low-resolution partial input. Stutz and Geiger [47] proposed an efficient weakly supervised framework to complete 3D shapes without slow optimization and direct supervision. These approaches often generate 3D shapes from random vectors and partial shapes and they cannot control the generated shapes precisely. In contrast to these approaches, ours can control the created 3D shapes in detail with the assistance of conceptual knowledge.

3 METHOD

3.1 Overview

Our framework is designed to explicitly combine graph-based conceptual knowledge of the object category with deep neural networks for 3D shape reconstruction from a single RGB image. Following previous works, we represent the 3D shape with 3D occupancy volume, where the value

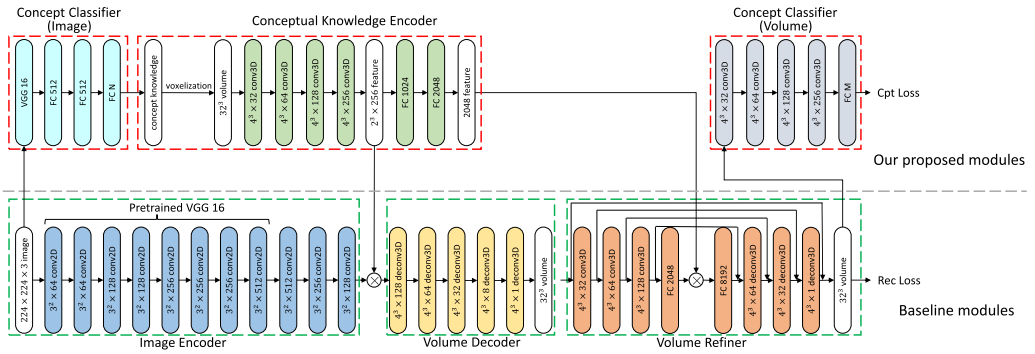

Fig. 2. Framework of our proposed approach. The framework consists of six modules: *concept classifier (image)*, *image encoder*, *conceptual knowledge encoder*, *volume decoder*, *volume refiner*, and *concept classifier (volume)*. It takes an image as input and outputs a reconstructed 3D shape.

of each voxel in the volume is either 1 (occupied) or 0 (empty). We employ the state-of-the-art architecture, that is, Pix2Vox-A [67], and augment it with the proposed conceptual knowledge encoder branch and two concept classifier branches on it. The structure of our framework is shown in Figure 2. The proposed framework takes a single RGB image as input. The concept classifier (image) predicts conceptual knowledge (i.e., object category and conceptual part labels) from the input image. The image encoder and conceptual knowledge encoder extract image features and concept features from the input RGB image and the predicted conceptual knowledge, respectively. The concept features are then inserted into the volume decoder and volume refiner modules via feature concatenation to reconstruct a 3D shape. The concept classifier (volume) is used to extract conceptual part labels from the reconstructed 3D shape for model optimization.

3.2 Baseline Architecture

We choose Pix2Vox-A [67], a state-of-the-art method in the task of 3D volume reconstruction from a single RGB image, as the baseline architecture. In the single-view reconstruction configuration, Pix2Vox-A consists of three parts: image decoder, volume decoder, and volume refiner. The image encoder, which is composed of 2D convolutional layers, is used to extract a feature tensor of size $2 \times 2 \times 2 \times 2048$ from a $224 \times 224 \times 3$ RGB image. The 3D deconvolution-based volume decoder reconstructs 3D volume probability of size $32 \times 32 \times 32$. The refiner further refines the reconstructed 3D shape. It includes 3D convolutional layers, fully connected layers, and 3D deconvolutional layers, and outputs a 3D volume of size $32 \times 32 \times 32$. We refer readers to [67] for a detailed description.

3.3 Our Proposed Framework

To accomplish the concept-aware 3D reconstruction task, we propose three novel modules—*concept classifier (image)*, *conceptual knowledge encoder*, and *concept classifier (volume)*—and combine them with the baseline architecture. Note that the proposed modules are pipeline agnostic and can be easily applied to most 3D reconstruction pipelines.

3.3.1 Concept Classifier (Image). The image-based concept classifier is proposed to predict the object category and conceptual part labels of the object in the given input image. The object category and conceptual part labels consist of the conceptual knowledge of the image, which are used as input for the conceptual knowledge encoder module. The classifier consists of a VGG16 module and 3 fully connected layers. The channels of fully connected layers are 512, 512, and N, where N is defined by the numbers of object categories and part labels. The first two connected layers

are followed by leaky ReLU activation. The output of the image-based concept classifier is split into two parts: object category and conceptual part labels. The image-based concept classifier is pretrained using input images with ground-truth category labels and conceptual part labels. The loss functions are composed of a standard softmax cross entropy loss (for object category prediction) and a concept loss (for conceptual part label prediction). We fix the pretrained weights of this module in our concept-aware 3D reconstruction task.

3.3.2 Conceptual Knowledge Encoder. The conceptual knowledge encoder is proposed to extract features from input conceptual knowledge. The category concept represented by a knowledge graph is difficult to be directly integrated into the current reconstruction pipeline. Therefore, we represent the conceptual knowledge in the form of 3D prototype volumes, which are generated by calculating the mean shape of volumetric conceptual parts and composing them based on the concept graph. Take the Chair category as an example. We first calculate the mean shape of each conceptual part of all training data as prototype volumes. Then, for a chair with a conceptual knowledge graph of $\{back, seat, base\}$, we compose the prototype volumes of the existed parts (i.e., back, seat, and base) via voxel-wise union. The encoder aims to extract features from a $32 \times 32 \times 32$ input volume. It consists of four 3D convolutional layers with kernel size 4^3 and two fully connected layers. Each convolutional layer is followed by a batch normalization layer, a leaky ReLU activation, and a max pooling layer with kernel size 2^3. The channels of the convolutional layers are 32, 64, 128, and 256. Each fully connected layer is followed by a ReLU activation. The channels of the fully connected layers are 1024 and 2048. The encoder extracts two features from the input volume. The $2 \times 2 \times 2 \times 256$ feature volume from the last 3D convolutional layer is fed into the volume decoder, while the 2048-d feature vector from the last fully connected layer is fed into the volume refiner. In contrast to a volume decoder, the volume refiner takes in higher-level features, as it is a refining process based on the coarse shapes reconstructed by the volume decoder. We validated the effectiveness of the structure of our conceptual knowledge encoder in preliminary experiments.

3.3.3 Concept Classifier (Volume). The volume based concept classifier is proposed to detect the conceptual parts from the reconstructed 3D shapes, which is later used to calculate the differences between the reconstructed 3D shape and the predicted conceptual knowledge for network training. The classifier is composed of four 3D convolutional layers with kernel size 4^3 and 1 fully connected layer. Each convolutional layer is followed by a batch normalization layer, a leaky ReLU activation, and a max pooling layer with kernel size 2^3. The channels of the convolutional layers are 32, 64, 128, and 256. The fully connected layer is followed by a sigmoid activation. It outputs a 1D feature with size M, where M is the number of conceptual parts defined by each object category. The volume-based concept classifier is pretrained using the ground-truth 3D shapes and the corresponding conceptual knowledge. We fix the pretrained weights of the volume-based concept classifier in our concept-aware 3D reconstruction task. The concept classifier is deployed during the training stage only and discarded during the inference stage.

3.4 Loss Functions

The proposed framework is trained with two loss functions: reconstruction loss and concept loss. Following existing 3D volume reconstruction approaches [10, 67], we choose the mean cross-entropy between reconstructed volume and ground-truth volume as our reconstruction loss. It is defined as

$$L_{rec} = \frac{1}{N} \sum_{i=1}^{N} \Big(p_i' log(p_i) + (1 - p_i') log(1 - p_i) \Big), \tag{1}$$

where N is the number of voxels. p'_i and p_i represent the ground-truth voxel value and the predicted voxel value, respectively. The reconstruction loss forces the network to generate a 3D shape similar to the ground truth at the voxel level.

For concept loss, we choose the mean cross-entropy between predicted concept labels and ground-truth concept labels of input conceptual knowledge, which is defined as

$$L_{cpt} = \frac{1}{M}\sum_{i=1}^{M}(s'_i log(s_i) + (1 - s'_i) log(1 - s_i)), \tag{2}$$

where M is the number of conceptual parts of the category and s'_i and s_i represent the existence of i-th conceptual part of the input conceptual knowledge and the reconstructed 3D shape, respectively. The concept loss encourages the network to generate a 3D shape similar to the input conceptual knowledge at the concept level.

The concept loss is discarded in category-agnostic training. This is because the object categories have different conceptual parts, which should not be trained together using concept loss. In category-specific training, the final loss is composed of two losses

$$L = L_{rec} + \lambda L_{cpt}, \tag{3}$$

where λ is weight of L_{cpt}. Based on our experiments, we set λ to 0.01.

4 EXPERIMENTS

4.1 Datasets

We evaluate our approach on both synthetic datasets and real datasets. Specifically, we use the following three datasets in our experiments.

ShapeNet [5] is the largest 3D model dataset, containing 55 object categories and more than 50,000 unique 3D models. Following [67], we use a subset of 13 major categories with more than 40,000 3D models. Each 3D model is rendered in 24 random views to get the corresponding synthetic RGB images. We evaluate both category-agnostic models and category-specific models of our approach on this dataset. **Pascal3D+** [66] contains real images of 12 object categories collected from Pascal and ImageNet datasets. Each category contains more than 1,000 images with 3D annotations. We use 7 object categories[1] that exist in ShapeNet and fine-tune both category-agnostic and category-specific models using pretrained weights from the ShapeNet dataset in our experiments. **Pix3D** [48] contains more than 10,000 real images from 9 categories. All images are annotated with 3D annotations. Following [67], we directly test on 2,894 images of the chair category with our model, which is trained on the ShapeNet chair category with each 3D model rendered from 60 views.

The conceptual annotations of the 3D objects are collected from PartNet [34]. The categories that do not have the corresponding annotations were manually annotated. The statistics of the conceptual parts of all object categories on ShapeNet are listed in Table 1.

4.2 Evaluation Metric

We use the Intersection-over-Union (IoU) metric to evaluate the reconstruction quality of our method, which is widely used in 3D object reconstruction tasks. Higher IoU value means better reconstruction quality. The IoU is defined as

$$\text{IoU} = \frac{\sum_{i,j,k} I(p_{(i,j,k)} > t) I(p'_{i,j,k})}{\sum_{i,j,k} I[I(p_{(i,j,k)} > t) + I(p'_{i,j,k})]}, \tag{4}$$

[1] airplane, boat, car, chair, dining table, sofa, and TV monitor.

Table 1. Statistics of the Conceptual Parts of All Object Categories in the ShapeNet Dataset

Category	Conceptual Parts (number)
airplane	body (3,936); wing (3,820); tail (3,793)
bench	base (1,745); seat (1,798); back (1,252); left arm (886); right arm (822)
cabinet	body (1,571); door (403)
car	wheels (7,121); body (7,158)
chair	base (6,152); seat (6,349); back (6,329); left arm (2,877); right arm (2,885)
display	base (789); screen (945)
lamp	base (2,045); unit (2,040)
speaker	base (235); body (1,597)
rifle	body (2,360); grip (2,003); magazine (1,607); stock(2,042)
sofa	seat (3,155); back (3,129); left arm (2,803); right arm (2,776)
table	base (8,012); top (8,012)
phone	body (1,050); lip (102)
vessel	body (1,937); building (1,681)

where $p_{i,j,k}$ and $p'_{i,j,k}$ represent the predicted probability and ground truth at (i, j, k), respectively. $I(\cdot)$ is an indicator function and t is the threshold to voxelize the probability. t is set to 0.4 in all experiments.

4.3 Implementation Details

In our proposed framework, the size of the input RGB image is $224 \times 224 \times 3$. The conceptual knowledge is represented as a 3D prototype volume of size 32^3 and the reconstructed 3D volume is 32^3 in size. For all datasets, we use the same configurations of Pix2Vox [67] for fair comparison. We implement our method using TensorFlow [1] and train the model with Adam optimizer [27] with β_1 and β_2 set to 0.9 and 0.999, respectively. For category-agnostic models, we train the network with a batch size of 64 for 300 epochs. The initial learning rate is 0.001 and decayed by 2 after 150 epochs. For category-specific models, we fine-tune another 200 epochs with a learning rate of 0.0005 from the pretrained category-agnostic models. We will release the code and processed dataset.

4.4 Shape Reconstruction from Image

4.4.1 Reconstruction on Synthetic Data. To evaluate the performance of our approach on synthetic data, we compare our approach on the ShapeNet dataset with the following state-of-the-art approaches: 3D-R2N2 [10], OGN [50], PSGN [13], Matryoshka [44], Voxel-Tube [44], Pix2Vox [67], and Pix2Vox++ [68]. Pix2Vox++ is an extension of Pix2Vox that replaces VGG16 with ResNet50 for image feature extraction and uses a better fusion module. Our backbone is VGG16 and is same as Pix2Vox. For comparison on category-agnostic models, we directly report the evaluation results from the original article to compare them with our approach. For comparison on category-specific models, we fine-tune category-specific models on each object category to compare our approach with Pix2Vox and Pix2Vox++ (Pix2Vox*, Pix2Vox++*, and Ours*). To validate the impact of the concept loss, we also train category-specific models without concept loss (Ours* w/o L_{cpt}) and evaluate the trained models.

Table 2 shows the IoU results of our method and the comparison methods on the ShapeNet dataset. Our approach outperforms the state-of-the-art approaches by a large margin. On category-agnostic models, our approach outperforms the corresponding baseline (i.e., Pix2Vox) in 11 out of 13 categories. The overall improvement is +1.5% using the IoU evaluation metric. Although

Table 2. Comparison of Single-View 3D Object Reconstruction on ShapeNet with Volume Size 32^3 Using IoU (in %) Evaluation Metric

	airplane	bench	cabinet	car	chair	display	lamp	speaker	rifle	sofa	table	telephone	vessel	all
3D-R2N2 [10]	51.3	42.1	71.6	79.8	46.6	46.8	38.1	66.2	54.4	62.8	51.3	66.1	51.3	56.0
OGN [50]	58.7	48.1	72.9	82.8	48.3	50.2	39.8	63.7	59.3	64.6	53.6	70.2	**63.2**	59.6
Matryoshka [44]	64.7	57.7	77.6	85.0	54.7	53.2	40.8	70.1	61.6	68.1	57.3	75.6	59.9	63.5
Voxel-Tube [44]	67.1	**63.7**	76.7	82.1	55.0	53.4	43.6	68.1	62.6	69.0	57.3	74.2	59.1	64.1
PSGN [13]	60.1	55.0	77.1	83.1	54.4	55.2	46.2	**73.7**	60.4	70.8	60.6	74.9	61.1	64.0
Pix2Vox [67]	68.4	61.6	79.2	85.4	56.7	53.7	44.3	71.4	61.5	70.9	60.1	77.6	59.4	66.1
Pix2Vox++ [68]	67.4	60.8	**79.9**	**85.8**	58.1	54.8	45.7	72.1	61.7	72.5	**62.0**	**80.9**	60.3	67.0
Ours	**69.2**	60.7	79.4	85.4	**59.2**	**56.0**	**47.4**	71.8	**63.6**	**72.9**	61.4	80.8	60.6	**67.6**
Pix2Vox* [67]	69.2	61.7	79.8	85.7	57.0	53.9	45.7	72.1	63.4	71.5	60.0	77.9	60.6	66.5
Pix2Vox++* [68]	69.8	63.0	79.9	**86.2**	58.4	55.0	46.6	72.4	64.8	72.8	61.5	80.5	62.2	67.6
Ours* w/o L_{cpt}	72.1	61.8	80.3	85.8	60.0	56.9	48.0	73.2	64.6	73.5	60.7	81.0	61.2	68.3
Ours*	**72.8**	**63.6**	**80.6**	86.1	**60.7**	**57.1**	**48.7**	**73.3**	**65.9**	**73.9**	**61.6**	**81.7**	**62.4**	**68.9**

The last four methods are category-specific models, while the others are category-agnostic models. Higher IoU value means better result. The bold numbers are the best results.

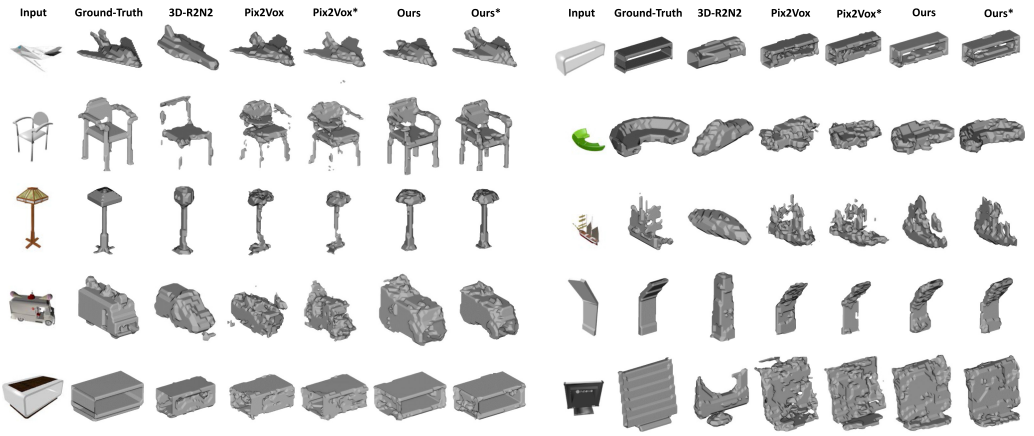

Fig. 3. Examples of the reconstruction results on ShapeNet with volume size 32^3. The reconstructed shapes are converted into 3D meshes for better visualization.

Pix2Vox++ adopted a more powerful feature extraction backbone (i.e., ResNet50), our approach still outperforms it by +0.6%. On category-specific models, the improvement of our approach is more obvious. Our approach outperforms Pix2Vox in all 13 categories. Our approach without concept loss (Ours* w/o L_{cpt}) and our final approach (Ours*) outperform Pix2Vox* by about +1.8% and +2.4%, respectively. Our final approach (Our*) outperforms Pix2Vox++* by +1.3%. We show some of the reconstructed samples in Figure 3. We can see that our approach not only reconstructs plausible 3D shapes on challenging data (car on left of the fourth row, and sofa on right of the second row) but also preserves details of reconstructed shapes (tail of airplane on left of the first row, and arms of chair on left of the second row).

To reconstruct a high-resolution 3D shape, we modify 3DensiNet [56], Pix2Vox [67], and our framework by adding 3D deconvolutional layers in the volume decoder and refiner. We also train

Table 3. Comparison of Single-View 3D Object Reconstruction on ShapeNet with Volume Size 64^3 Using IoU (in %) Evaluation Metric

	airplane	bench	cabinet	car	chair	display	lamp	speaker	rifle	sofa	table	telephone	vessel	all
3DensiNet [56]	54.6	38.5	72.1	79.7	45.5	43.6	33.3	67.3	47.1	64.5	49.6	75.3	50.8	55.5
Voxel-Tube [44]	60.2	46.3	74.4	81.6	51.3	**49.6**	36.9	68.2	50.9	67.4	52.9	75.4	53.5	59.1
Pix2Vox [67]	61.4	46.4	72.9	80.6	50.6	47.3	**37.9**	66.8	51.6	67.2	52.2	74.3	53.7	59.7
Ours	**63.6**	**48.5**	**75.8**	**82.1**	**52.9**	49.3	36.7	**68.5**	**52.5**	**68.9**	**54.0**	**77.1**	**54.8**	**61.4**
Pix2Vox* [67]	63.2	48.4	73.9	81.3	50.8	48.0	38.8	67.2	54.2	67.8	51.9	76.8	54.9	60.4
Ours*	**68.1**	**53.2**	**77.1**	**83.1**	**53.6**	**50.8**	**38.9**	**69.8**	**54.7**	**69.7**	**54.2**	**79.4**	**56.7**	**62.9**

The last two methods are category-specific models, while the others are category-agnostic models. Higher IoU value means better result. The bold numbers are the best results.

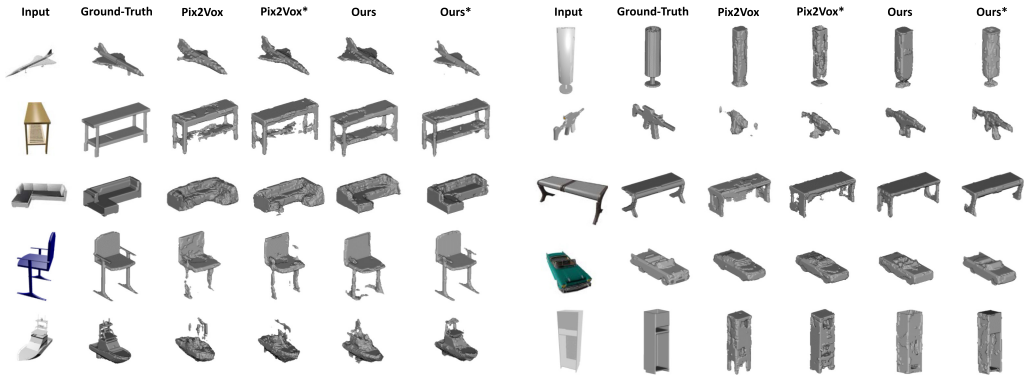

Fig. 4. Examples of the reconstruction results on ShapeNet with volume size 64^3. The reconstructed shapes are converted into 3D meshes for better visualization.

Voxel-Tube [44] with volume size 64^3 using the released code. The results are shown in Table 3. Our approach outperforms all comparison approaches. Compared with Pix2Vox, our approach outperforms it on both category-agnostic models and category-specific models. The overall improvements are +1.7% and +2.5%, respectively. The visualized results are in Figure 4. Our approach preserves more details of the reconstructed 3D shapes. For example, our approach successfully reconstructs the legs and arms of the chair (left of the fourth row), while Pix2Vox failed to generate reasonable shapes of these parts.

4.4.2 Reconstruction on Real-World Data. We compare our approach with previous approaches on Pascal3D+ and Pix3D datasets. Images in these datasets are all real images rather than synthetic ones. Quantitative comparison on Pascal3D+ is in Table 4. Our approach outperforms all previous approaches in all categories. Compared with Pix2Vox, our approach outperforms it on both category-agnostic and category-specific models by about +7.3% and +4.9%, respectively. Compared with results on ShapeNet, the reconstruction results are much better. This is because, in the Pascal3D+ dataset, some images share the same ground-truth 3D model, making the reconstruction much easier. On the Pix3D dataset, we predict the silhouette of the object in the input image using BlendMask [6] and mask the background with the predicted silhouette. The processed images are then fed into the pretrained model to reconstruct a 3D shape. In Table 5, we can see

Table 4. Comparison of Single-view 3D Object Reconstruction on Pascal3D+ With Volume Size 32^3 Using IoU (In %) Evaluation Metric

	aeroplane	boat	car	chair	diningtable	sofa	tvmonitor	all
3D-R2N2 [10]	54.4	56.0	69.9	28.0	-	33.2	57.4	-
DCT [22]	55.5	52.3	63.5	25.0	-	46.2	54.9	-
3DensiNet [56]	-	32.6	60.7	25.9	-	57.4	60.6	-
Pix2Vox [67]	60.4	67.3	79.4	41.1	39.8	66.6	54.9	58.5
Ours	**65.7**	**73.6**	**81.0**	**54.8**	**45.7**	**75.9**	**63.8**	**65.8**
Pix2Vox* [67]	71.2	76.3	82.3	57.8	46.3	80.6	64.7	68.5
Ours*	**75.6**	**84.7**	**85.1**	**64.0**	**46.5**	**84.3**	**73.3**	**73.4**

The last two methods are category-specific models, while the others are category-agnostic models. Higher IoU value means better result. The bold numbers are the best results.

Table 5. Comparison of Single-view 3D Object Reconstruction on Pix3D With Volume Size 32^3 Using IoU (In %) Evaluation Metric

Methods	3D-R2N2 [10]	3D-VAE-GAN [62]	MarrNet [61]	DAREC [42]	DRC [52]	Pix3D [48]	ShapeHD [63]	Pix2Vox [67]	Ours
IoU	13.6	17.1	23.1	24.1	26.5	28.2	28.4	28.8	**30.5**

Higher IoU value means better result. The bold numbers are the best results.

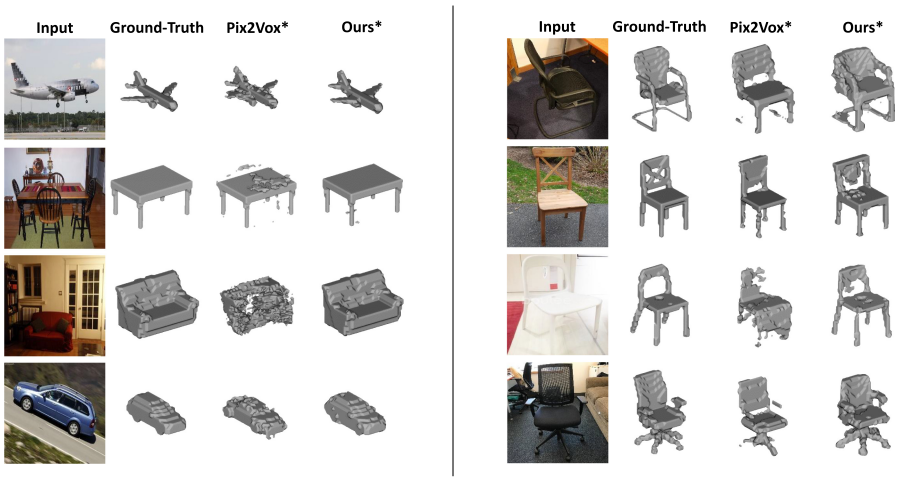

Fig. 5. Examples of the reconstruction results on Pascal3D+ (left) and Pix3D (right) datasets. The volume size is 32^3. The reconstructed shapes are converted into 3D meshes for better visualization.

that our approach outperforms all previous approaches. Compared with Pix2Vox, our approach outperforms it by about +1.7%. In Figure 5, we show examples of reconstructed shapes on real images. Our approach can reconstruct accurate 3D shapes on challenging data (left of the third row, where the sofa object is similar to the background), while Pix2Vox cannot even reconstruct a sofa-like shape. The reconstructed 3D shapes of our approach are also topologically consistent with the images. For example, our approach reconstructs a chair with a hole on its back based on the corresponding chair image (right image on the third row).

Table 6. Comparison of Single-View 3D Object Reconstruction Using Different Reconstruction Pipelines on ShapeNet with Volume Size 32^3 Using IoU (in %) Evaluation Metric

	airplane	bench	cabinet	car	chair	display	lamp	speaker	rifle	sofa	table	telephone	vessel	all
3DensiNet [56]	-	-	74.8	81.3	46.5	48.7	-	-	-	66.8	54.5	67.4	55.4	61.9
3DensiNet cpt	-	-	**75.1**	**83.2**	**52.0**	**50.0**	-	-	-	**68.0**	**56.0**	**77.6**	**55.8**	**64.7**
Voxel-Tube [44]	**67.1**	**63.7**	76.7	82.1	55.0	53.4	43.6	68.1	**62.6**	69.0	57.3	74.2	59.9	64.1
Voxel-Tube cpt	66.0	58.1	**77.6**	**84.3**	**57.6**	**56.3**	**47.8**	**71.9**	61.5	**71.0**	**59.0**	**79.0**	**60.0**	**65.9**
Pix2Vox [67]	68.4	**61.6**	79.2	85.4	56.7	53.7	44.3	71.4	61.5	70.9	60.1	77.6	59.4	66.1
Pix2Vox cpt	**69.2**	60.7	**79.4**	85.4	**59.2**	**56.0**	**47.4**	**71.8**	**63.6**	**72.9**	**61.4**	**80.8**	**60.6**	**67.6**

Higher IoU value means better result. The bold numbers are the best results.

4.5 Ablation Study

In this section, we validate the effectiveness of our pipeline-agnostic conceptual knowledge framework and our prototype volume-based conceptual knowledge representation.

4.5.1 Effectiveness of Our Conceptual Knowledge Framework. Since our conceptual knowledge framework is pipeline agnostic, we select some representative vanilla 3D reconstruction pipelines (i.e., 3DensiNet [56], Voxel-Tube [44], and Pix2Vox [67]) and apply our conceptual knowledge framework on them (namely, 3DensiNet cpt, Voxel-Tube cpt, and Pix2Vox cpt). For all pipelines, encoded conceptual features are concatenated with encoded image features before the volume decoder/refiner module. We use the same configurations with these pipelines for fair comparison. In Table 6, we can see that our conceptual knowledge framework enhances all pipelines with a good margin (from +1.5% to +2.8%), proving the effectiveness of our conceptual knowledge framework.

4.5.2 Effectiveness of Our Prototype Volume-Based Conceptual Knowledge Representation. To validate the effectiveness of our prototype volume-based conceptual knowledge representation, we compare it with a baseline in which the conceptual knowledge is represented with semantic labels. That is, the conceptual knowledge is represented with object category labels and object part labels. Take the Chair category in the ShapeNet dataset as an example. This category is represented with a one-hot 13-D vector (13 object categories) and a 5-D vector (Chair is composed of 5 semantic parts). We use a 6-layer MLP network to extract conceptual features from the semantic representation vectors and concatenate them with image features for 3D shape reconstruction. Pix2Vox is selected as the baseline pipeline in this experiment. The comparison results are in Table 7. We can see that compared with semantic label-based conceptual knowledge representation, our prototype volume-based conceptual knowledge representation enhances the original pipeline with a good margin (+1.5% vs. +0.7%). This is because our prototype volume-based conceptual knowledge representation involves more detailed information of object concepts, such as coarse shapes of different parts and relative locations between object parts. Such concrete information can better guide the reconstruction network than the abstract semantic information.

4.6 Additional Applications

Since our approach can utilize conceptual knowledge along with an image as input, it can be directly applied in some other tasks. Here, we present two possible applications with the proposed framework.

4.6.1 Concept-Assisted Shape Creation. By modifying the predicted conceptual knowledge, we can create novel 3D shapes from a single image or multiple images. Given an image, we can create

Table 7. Comparison of Different Concept Knowledge Representations on ShapeNet With Volume Size 32^3 Using IoU (in %) Evaluation Metric

	airplane	bench	cabinet	car	chair	display	lamp	speaker	rifle	sofa	table	telephone	vessel	all
baseline	68.4	**61.6**	79.2	85.4	56.7	53.7	44.3	71.4	61.5	70.9	60.1	77.6	59.4	66.1
baseline + semantic label	68.5	60.8	78.6	85.0	58.1	55.0	**47.5**	71.7	63.5	72.3	59.5	**82.6**	60.5	66.8
baseline + prototype volume (ours)	**69.2**	60.7	**79.4**	85.4	**59.2**	**56.0**	47.4	**71.8**	**63.6**	**72.9**	**61.4**	80.8	**60.6**	**67.6**

Higher IoU value means better result. The bold numbers are the best results.

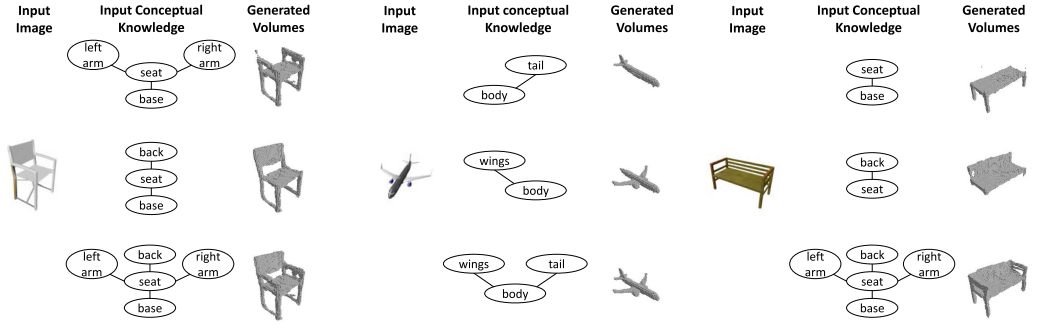

Fig. 6. Creating various 3D shapes of input images with different conceptual knowledge.

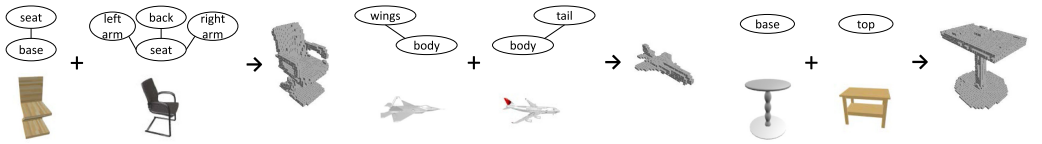

Fig. 7. Creating novel shapes by composing shapes from multiple images.

different 3D shapes using various conceptual knowledge. For example, given an image of a chair, we can remove the arms in predicted concepts and feed the graph {*back*, *seat*, *base*} to the network. The reconstructed shape is a 3D chair without arms. In Figure 6, we show some of the generated 3D shapes with different types of input conceptual knowledge. The generated 3D shapes satisfy the following requirements: (1) the 3D shapes are consistent with input images on the geometry level and (2) the 3D shapes are consistent with input knowledge on the concept level.

We can also create novel shapes from multiple images. First, we decompose the conceptual knowledge graph of the object into different sub-graphs. Then, for each sub-graph, we generate a 3D shape with our approach using one image and the sub-graph as inputs. All generated shapes are finally composed through the voxel-wise union to create the desired novel 3D object. In Figure 7, we present the created 3D shapes from multiple images. Consider chair composition as an example. We decompose the conceptual knowledge graph of a chair into {seat, base} and {back, seat, left arm, right arm}. Then, we generate one 3D shape using one chair image with a {seat, base} graph and one 3D shape using another chair image with a {back, seat, left arm, right arm} graph. Finally, a voxel-wise union operation is used to compose the two generated shapes into a novel 3D shape.

4.6.2 Concept-Assisted Semantic Shape Reconstruction. Our approach can be used for the semantic 3D shape reconstruction task by treating each concept as a semantic part. To do semantic reconstruction, we first reconstruct 3D shapes of an image with various input conceptual

Table 8. Comparison of Semantic Reconstruction on ShapeNet (Chair, Table, and Display) with Volume Size 32^3 Using IoU (in %) Evaluation Metric

	chair			table	display	
	base	arms	back	top	base	screen
Baseline	27.0	20.7	34.7	39.6	48.4	34.6
Shi [45]	39.0	**30.8**	46.1	44.7	48.0	38.4
AICNet [30]	25.1	25.1	39.4	58.4	25.1	**42.7**
Ours	**56.4**	27.1	**47.8**	**70.9**	**57.9**	33.5

Higher IoU value means better result. The bold numbers are the best results.

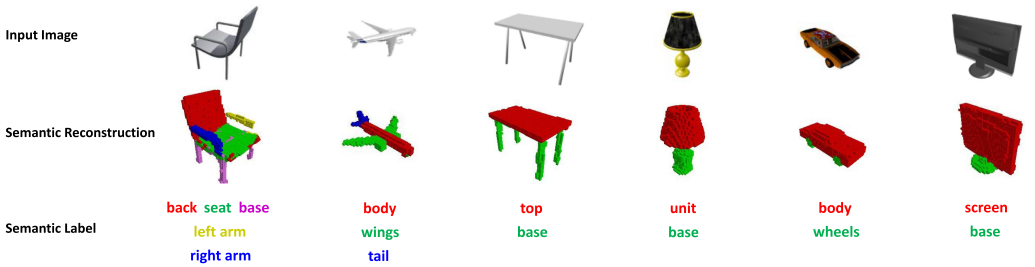

Fig. 8. Reconstructing semantic 3D shapes via conceptual knowledge–based shape reconstruction and subtraction.

knowledge. The semantic parts of the 3D shape are then acquired by 3D shape subtraction. For example, a chair is composed of base, seat, arms, and back. To get the semantic shape of the back, we can subtract the generated shape using {base, seat} knowledge from the generated shape using {base, seat, back}. Other semantic parts are acquired similarly. Here, we quantitatively compare the semantic reconstruction results with three approaches. (1) Baseline: an encoder-decoder architecture that is trained end-to-end to reconstruct 3D part volumes from input image. We directly use the results of the baseline model from [45]. (2) Shi's work [45]: A state-of-the-art part-level reconstruction approach that adopts a local awareness module to assist the semantic reconstruction task. (3) AICNet [30]: A state-of-the-art semantic scene reconstruction approach to reconstruct semantic volumes from depth and RGB images. We modify the network by replacing 2D feature extraction and downsample modules with 3D feature extraction modules so that it can be used to obtain 3D semantic volumes from reconstructed 3D shapes. In Table 8, we show the comparison results of common categories and object parts using the IoU evaluation metric. We can see that our approach outperforms Shi's on parts with larger shapes, such as the base and back of the chair and top of the table. However, on smaller parts (i.e., the arms of the chair), Shi's approach works better. Compared with AICNet, our approach outperforms it on all parts except the display screen. Note that we directly apply our approach on this semantic reconstruction task without any fine-tuning or retraining. Visualization of shapes generated with our approach are in Figure 8. We can see that our approach can reconstruct the 3D shapes as well as classify the semantic labels of each voxel accurately.

4.7 Discussion and Limitations

By adding the three modules—that is, concept classifier (image), conceptual knowledge encoder, and concept classifier (volume)—on the Pix2Vox baseline architecture, the reconstruction

performance improves on all evaluation datasets (+1.5% at least and +7.3% at most). Here, we discuss the impact of conceptual knowledge in this task.

Previous 3D reconstruction approaches reconstruct 3D shapes in a bottom-up way, which directly predicts probability values of voxels from encoded image features. The loss function penalizes only the dissimilarity of predicted voxels and ground-truth voxels. Compared with previous approaches, our approach has the following benefits:

- Our use of conceptual knowledge provides a global understanding of the objects and plays a dominant role in the task of 3D shape reconstruction. The concept loss also forces the network to pay more attention to the structure of the objects and learn high-level features at the training stage.
- The volume decoder and refiner reconstructs 3D shapes from fused features, which combines semantic attributes (from conceptual knowledge) and vision features (from the image). Compared with 3D reconstruction from single modal features, reconstructing 3D shapes from the fused multi-modal features can be much easier and more effective.
- With differing conceptual knowledge as input, one image can be related to multiple ground-truth 3D shapes, which inherently acts as data augmentation to increase the generality of the network and reduce the over-fitting problem.

Our approach also has some limitations. Since we added three novel modules containing 3D convolution layers, our approach contains more parameters compared with the baseline (161M vs. 114M), which leads to a longer training time (45 h vs. 25 h). Although we can reconstruct high-resolution 3D shapes by adding 3D convolutional and deconvolutional layers in the reconstruction pipeline, the training becomes much slower and the parameters are much larger because of the increased memory of the volume representation of 3D shapes. In the future, we plan to use memory-efficient 3D shape representations, such as octrees and point clouds, to deal with these limitations.

5 CONCLUSION

In this article, we propose a novel multimodal framework to explicitly combine graph-based conceptual knowledge with deep neural networks for accurate 3D shape reconstruction from a single RGB image. Experiments on three benchmark datasets (i.e., ShapeNet, Pascal3D+, and Pix3D) prove that our approach outperforms state-of-the-art methods. In addition to reconstructing 3D shapes from images, our approach can also create various novel 3D shapes with the assistance of conceptual knowledge, which can be used in 3D shape design and editing. In future work, we plan to extend our approach to reconstruct 3D shapes with other shape representations, such as point clouds, mesh, and octrees, for efficient high-resolution 3D shape reconstruction.

ACKNOWLEDGMENT

We would like to thank Ce Zhu, Shuai Liu, and Jialei Ai for running and generating samples of comparison methods. We also thank all of the volunteers who significantly helped us with the conceptual annotation work of the 3D objects.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. tensorflow: A system for large-scale machine learning. In *USENIX Symposium on Operating Systems Design and Implementation*. 265–283.

[2] Hassan Afzal, Djamila Aouada, Bruno Mirbach, and Björn Ottersten. 2018. Full 3D reconstruction of non-rigidly deforming objects. *ACM Transactions on Multimedia Computing, Communications, and Applications* 14, 1s (2018), 1–23.

[3] Amir Arsalan Soltani, Haibin Huang, Jiajun Wu, Tejas D. Kulkarni, and Joshua B. Tenenbaum. 2017. Synthesizing 3D shapes via modeling multi-view depth maps and silhouettes with deep generative networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu. IEEE. 1511–1519.

[4] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J. Leonard. 2016. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics* 32, 6 (2016), 1309–1332.

[5] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. 2015. shapenet: An information-rich 3D model repository. *arXiv:1512.03012*.

[6] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. 2020. Blendmask: Top-down meets bottom-up for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Virtual. IEEE. 8573–8581.

[7] Zhiqin Chen and Hao Zhang. 2019. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Long Beach. IEEE. 5939–5948.

[8] K. M. G. Cheung, Simon Baker, and Takeo Kanade. 2003. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1, Wisconsin. IEEE. 1–1.

[9] Seonghwa Choi, Anh-Duc Nguyen, Jinwoo Kim, Sewoong Ahn, and Sanghoon Lee. 2019. Point cloud deformation for single image 3d reconstruction. In *IEEE International Conference on Image Processing (ICIP'19)*. Taipei. IEEE. 2379–2383.

[10] Christopher B. Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 2016. 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In *Proceedings of the European Conference on Computer Vision (Lecture Notes in Computer Science)*, Vol. 9912, Springer. Amsterdam. Springer. 628–644.

[11] Angela Dai, Charles Ruizhongtai Qi, and Matthias Nießner. 2017. Shape completion using 3D-encoder-predictor CNNs and shape synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu. IEEE. 5868–5877.

[12] Hugh Durrant-Whyte and Tim Bailey. 2006. Simultaneous localization and mapping. *IEEE Robotics & Automation Magazine* 13, 2 (2006), 99–110.

[13] Haoqiang Fan, Hao Su, and Leonidas J. Guibas. 2017. A point set generation network for 3D object reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu. IEEE. 605–613.

[14] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. 2019. Mesh R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*. Long Beach. IEEE. 9785–9795.

[15] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan C. Russell, and Mathieu Aubry. 2018. A papier-mâché approach to learning 3D surface generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City. IEEE. 216–224.

[16] Xiaoguang Han, Zhen Li, Haibin Huang, Evangelos Kalogerakis, and Yizhou Yu. 2017. High-resolution shape completion using deep neural networks for global structure and local geometry inference. In *Proceedings of the IEEE International Conference on Computer Vision*. Honolulu. IEEE. 85–93.

[17] Xian-Feng Han, Hamid Laga, and Mohammed Bennamoun. 2019. Image-based 3D object reconstruction: State-of-the-art and trends in the deep learning era. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 5 (2019), 1578–1604.

[18] Christian Häne, Shubham Tulsiani, and Jitendra Malik. 2017. Hierarchical surface prediction for 3D object reconstruction. In *International Conference on 3D Vision*. Qingdao. IEEE. 412–420.

[19] Richard Hartley and Andrew Zisserman. 2003. *Multiple View Geometry in Computer Vision*. Cambridge University Press.

[20] Eldar Insafutdinov and Alexey Dosovitskiy. 2018. Unsupervised learning of shape and pose with differentiable point clouds. In *Advances in Neural Information Processing Systems*. Montréal. MIT Press. 2802–2812.

[21] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. 2011. KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*. Santa Barbara. ACM. 559–568.

[22] Adrian Johnston, Ravi Garg, Gustavo Carneiro, Ian Reid, and Anton van den Hengel. 2017. Scaling CNNs for high resolution volumetric reconstruction from a single image. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. Venice. IEEE. 939–948.

[23] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. 2018. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (Lecture Notes in Computer Science)*, Vol. 11219, *Munich*. Springer. 386–402.

[24] Abhishek Kar, Christian Häne, and Jitendra Malik. 2017. Learning a multi-view stereo machine. In *Advances in Neural Information Processing Systems*. Long Beach. MIT Press, 365–376.

[25] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. 2015. Category-specific object reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston. IEEE. 1966–1974.

[26] Hiroharu Kato and Tatsuya Harada. 2019. Learning view priors for single-view 3D reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Long Beach. IEEE. 9778–9787.

[27] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980* (2014).

[28] Andrey Kurenkov, Jingwei Ji, Animesh Garg, Viraj Mehta, Junyoung Gwak, Christopher Choy, and Silvio Savarese. 2018. Deformnet: Free-form deformation network for 3D shape reconstruction from a single image. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. Lake Tahoe. IEEE. 858–866.

[29] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.

[30] Jie Li, Kai Han, Peng Wang, Yu Liu, and Xia Yuan. 2020. Anisotropic convolutional networks for 3D semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Virtual. IEEE. 3351–3359.

[31] Yi-Lun Liao, Yao-Cheng Yang, Yuan-Fang Lin, Pin-Jung Chen, Chia-Wen Kuo, Wei-Chen Chiu, and Yu-Chiang Frank Wang. 2019. Learning pose-aware 3D reconstruction via 2D-3D self-consistency. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. Brighton. IEEE. 3857–3861.

[32] Chen-Hsuan Lin, Chen Kong, and Simon Lucey. 2018. Learning efficient point cloud generation for dense 3D object reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32, New Orleans. AAAI, 1–1.

[33] Donald J. R. Meagher. 1980. *Octree Encoding: A New Technique for the Representation, Manipulation and Display of Arbitrary 3D Objects by Computer*. Electrical and Systems Engineering Department, Rensseiaer Polytechnic.

[34] Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. 2019. partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Long Beach. IEEE. 909–918.

[35] K. L. Navaneet, Priyanka Mandikal, Mayank Agarwal, and R. Venkatesh Babu. 2019. capnet: Continuous approximation projection for 3D point cloud reconstruction using 2D supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, Honolulu. AAAI. 8819–8826.

[36] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. 2011. KinectFusion: Real-time dense surface mapping and tracking. In *IEEE International Symposium on Mixed and Augmented Reality*. Basel. IEEE. 127–136.

[37] Anh-Duc Nguyen, Seonghwa Choi, Woojae Kim, and Sanghoon Lee. 2019. Graphx-convolution for point cloud deformation in 2D-to-3D conversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Seoul. IEEE. 8628–8637.

[38] Pietro Pala and Stefano Berretti. 2019. Reconstructing 3D face models by incremental aggregation and refinement of depth frames. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15, 1 (2019), 1–24.

[39] Junyi Pan, Xiaoguang Han, Weikai Chen, Jiapeng Tang, and Kui Jia. 2019. Deep mesh reconstruction from single RGB images via topology modification networks. In *Proceedings of the IEEE International Conference on Computer Vision*. Seoul. IEEE. 9964–9973.

[40] Junyi Pan, Jun Li, Xiaoguang Han, and Kui Jia. 2018. Residual meshnet: Learning to deform meshes for single-view 3D reconstruction. In *International Conference on 3D Vision*. Verona. IEEE. 719–727.

[41] Yun-he Pan. 2019. On visual knowledge. *Frontiers of Information Technology & Electronic Engineering* 20, 8 (2019), 1021–1025.

[42] Pedro O. Pinheiro, Negar Rostamzadeh, and Sungjin Ahn. 2019. Domain-adaptive single-view 3D reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Seoul. IEEE. 7638–7647.

[43] Stephan R. Richter and Stefan Roth. 2015. Discriminative shape from shading in uncalibrated illumination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston. IEEE. 1128–1136.

[44] Stephan R. Richter and Stefan Roth. 2018. Matryoshka networks: Predicting 3D geometry via nested shape layers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City. IEEE. 1936–1944.

[45] Dingfeng Shi, Yifan Zhao, and Jia Li. 2020. Reconstructing part-level 3D models from a single image. In *2020 IEEE International Conference on Multimedia and Expo*. Virtual. IEEE. 1–6.

[46] Edward Smith, Scott Fujimoto, Adriana Romero, and David Meger. 2019. GEOMetrics: Exploiting geometric structure for graph-encoded objects. In *Proceedings of the International Conference on Machine Learning*, Vol. 97, Long Beach. ACM. 5866–5876.

[47] David Stutz and Andreas Geiger. 2020. Learning 3D shape completion under weak supervision. *International Journal of Computer Vision* 128, 5 (2020), 1162–1181.

[48] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B. Tenenbaum, and William T. Freeman. 2018. Pix3D: Dataset and methods for single-image 3D shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City. IEEE. 2974–2983.

[49] Jiapeng Tang, Xiaoguang Han, Junyi Pan, Kui Jia, and Xin Tong. 2019. A skeleton-bridged deep learning approach for generating meshes of complex topologies from single RGB images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Long Beach. IEEE. 4541–4550.

[50] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. 2017. Octree generating networks: Efficient convolutional architectures for high-resolution 3D outputs. In *Proceedings of the IEEE International Conference on Computer Vision*. Venice. IEEE. 2088–2096.

[51] Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. 2018. Multi-view consistency as supervisory signal for learning shape and pose prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City. IEEE. 2897–2905.

[52] Shubham Tulsiani, Tinghui Zhou, Alexei A. Efros, and Jitendra Malik. 2017. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2626–2634.

[53] Hanqing Wang, Jiaolong Yang, Wei Liang, and Xin Tong. 2019. Deep single-view 3D object reconstruction with visual hull embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, Honolulu. AAAI, 8941–8948.

[54] Jianren Wang and Zhaoyuan Fang. 2020. GSIR: Generalizable 3D shape interpretation and reconstruction. In *European Conference on Computer Vision*. Virtual. Springer. 498–514.

[55] Jinglu Wang, Bo Sun, and Yan Lu. 2019. MVPNet: Multi-view point regression networks for 3D object reconstruction from a single image. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, Honolulu. AAAI, 8949–8956.

[56] Meng Wang, Lingjing Wang, and Yi Fang. 2017. 3DensiNet: A robust neural network architecture towards 3D volumetric object prediction from 2D image. In *Proceedings of the ACM International Conference on Multimedia*. Mountain View. ACM. 961–969.

[57] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Hang Yu, Wei Liu, Xiangyang Xue, and Yu-Gang Jiang. 2020. Pixel2Mesh: 3D mesh model generation via image guided deformation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[58] Weiyue Wang, Qiangui Huang, Suya You, Chao Yang, and Ulrich Neumann. 2017. Shape inpainting using 3D generative adversarial network and recurrent convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*. Venice. IEEE. 2298–2306.

[59] Matthew J. Westoby, James Brasington, Niel F. Glasser, Michael J. Hambrey, and Jennifer M. Reynolds. 2012. 'structure-from-motion' photogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology* 179 (2012), 300–314.

[60] Andrew P. Witkin. 1981. Recovering surface shape and orientation from texture. *Artificial Intelligence* 17, 1-3 (1981), 17–45.

[61] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum. 2017. MarrNet: 3D shape reconstruction via 2.5D sketches. In *Advances in Neural Information Processing Systems*, Long Beach. MIT Press, 540–550.

[62] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. 2016. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In *Advances in Neural Information Processing Systems*. 82–90.

[63] Jiajun Wu, Chengkai Zhang, Xiuming Zhang, Zhoutong Zhang, William T. Freeman, and Joshua B. Tenenbaum. 2018. Learning shape priors for single-view 3D completion and reconstruction. In *Proceedings of the European Conference on Computer Vision (Lecture Notes in Computer Science)*, Vol. 11215, Springer. 673–691.

[64] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 2015. 3D ShapeNets: A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston. IEEE. 1912–1920.

[65] Nan Xiang, Li Wang, Tao Jiang, Yanran Li, Xiaosong Yang, and Jianjun Zhang. 2019. Single-image mesh reconstruction and pose estimation via generative normal map. In *Proceedings of the International Conference on Computer Animation and Social Agents*. Paris. ACM. 79–84.

[66] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. 2014. Beyond Pascal: A benchmark for 3D object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision*. Steamboat Springs. IEEE. 75–82.

[67] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. 2019. Pix2Vox: Context-aware 3D reconstruction from single and multi-view images. In *Proceedings of the IEEE International Conference on Computer Vision*. Seoul. IEEE. 2690–2698.

[68] Haozhe Xie, Hongxun Yao, Shengping Zhang, Shangchen Zhou, and Wenxiu Sun. 2020. Pix2Vox++: Multi-scale context-aware 3D object reconstruction from single and multiple images. *International Journal of Computer Vision* 128, 12 (2020), 2919–2935.

[69] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. 2016. Perspective transformer nets: Learning single-view 3D object reconstruction without 3D supervision. In *Advances in Neural Information Processing Systems*, Barcelona. MIT Press, 1696–1704.

[70] Bo Yang, Stefano Rosa, Andrew Markham, Niki Trigoni, and Hongkai Wen. 2018. Dense 3D object reconstruction from a single depth view. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 12 (2018), 2820–2834.

[71] Guandao Yang, Yin Cui, Serge Belongie, and Bharath Hariharan. 2018. Learning single-view 3D reconstruction with limited pose supervision. In *Proceedings of the European Conference on Computer Vision (Lecture Notes in Computer Science)*, Vol. 11219, Munich. Springer, 90–105.

[72] Shuo Yang, Min Xu, Haozhe Xie, Stuart Perry, and Jiahao Xia. 2021. Single-view 3D object reconstruction from shape priors in memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE. 3152–3161.

[73] Chun-Han Yao, Wei-Chih Hung, Varun Jampani, and Ming-Hsuan Yang. 2021. Discovering 3D parts from image collections. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE. 12981–12990.

[74] Yuan Yao, Nico Schertler, Enrique Rosales, Helge Rhodin, Leonid Sigal, and Alla Sheffer. 2020. start here Front2Back: Single view 3D shape reconstruction via front to back prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Virtual. IEEE. 531–540.

[75] Chen Zhang. 2019. CuFusion2: Accurate and denoised volumetric 3D object reconstruction using depth cameras. *IEEE Access* 7 (2019), 49882–49893.

[76] Chen Zhang and Yu Hu. 2017. CuFusion: Accurate real-time camera tracking and volumetric scene reconstruction with a cuboid. *Sensors* 17, 10 (2017), 2260.

[77] Qian-Yi Zhou and Vladlen Koltun. 2015. Depth camera tracking with contour cues. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston. IEEE. 632–638.

Received December 2021; revised August 2021; accepted September 2021