Structure-Aware 3D Shape Synthesis from Single-View Images

Xuyang Hu[1]
huxuyangahu@163.com
Fan Zhu[2]
fan.zhu@inceptioniai.org
Li Liu[2]
liuli1213@gmail.com
Jin Xie[3]
csjxie@njust.edu.cn
Jun Tang[1]
tangjunahu@163.com
Nian Wang[1]
wn_xlb@ahu.edu.cn
Fumin Shen[4]
fumin.shen@gmail.com
Ling Shao[2]
ling.shao@ieee.org

[1] Anhui University, China
[2] Inception Institute of Artificial Intelligence
[3] Nanjing University of Science and Technology
[4] University of Electronic Science and Technology of China

Abstract

Automated 3D prototyping capability is at critical demands for rapid production across multiple industries. While traditional 3D object reconstruction approaches have been heavily relying on depth maps, which are either costly to acquire or inaccurate to approximate. Some recent approaches attempted to synthesize 3D shapes from monocular clues by directly learning complex non-linear transformations for bridging the cross-domain data. Despite that visually satisfactory synthesized 3D objects can be observed, these 3D synthesis approaches have following major restrictions: 1) the learned non-linear transformations are hardly aware of the intra-class objects' structural variations; 2) multiple-image inputs from different observation viewpoints are required for generating structure-aware 3D shapes; 3) objects are always observed from natural images with cluttered backgrounds. In this work, we aim to address above restrictions by proposing an improved 3D shape synthesis method that relies on a single input image. Benefiting from recent advancements of generative models, we learn to map the distributions between different-view 2D images and eventually generate multi-view images from a single image. The generated multi-view images are then employed to synthesize 3D shapes with incremental object details. In addition, by building perfectly aligned object poses in cross-view images as well as the corresponding 3D shape, the 2D-to-3D mapping can be guided to be aware of geometric structure of the objects. Extensive quantitative and qualitative results demonstrate that the proposed approach can achieve the state-of-the-art performance on the ShapeNet datasets.

1 Introduction

3D shape reconstruction is a classical problem in 3D computer vision. In the recent decade, along with the exponentially growing industrial needs that range from virtual reality to 3D printing, the capability for automated 3D prototyping becomes a critical technique that can potentially promote the development of several industries. Correspondingly, 3D object reconstruction has been receiving increasingly extensive attentions as a research scope [4, 11, 15, 34]. Among the popular existing 3D object reconstruction approaches, the depth estimation [20, 30] and the monocular cues are the commonly used for inferring 3D shapes. Considering the additional costs and complex procedures led by external hardware equipment for acquiring the depth maps, depth estimation from natural images can be one of the most straightforward solutions for recovering 3D objects. However, despite that deep learning has recently demonstrated some progress on the depth estimation task [5, 19, 20], inferring the depth channel from single images is arguably an ill-posed problem due to the missing partial information on unobserved side views of the objects [6]. Compare to depth map estimation, recovering the full 3D shape from images is a more challenging task due to the requirement of recovering detailed shape poses and structures. When utilizing monocular cues (e.g., silhouettes, shading and texture) for full 3D shape reconstruction, the state-of-the-art approaches [4, 14, 31] share the following common restrictions, such as the requirement of multiple image captures from dense viewpoints and non-reflective objects' appearances. Benefiting from the tremendous success of convolutional neural network (CNN) [17], the advancement of deep generative models and the continuing growing 3D shape repositories, more possibilities have been explored for the generation of 3D shapes from 2D images. A typical example is the 3D Recurrent Reconstruction Neural Network (3D-R2N2) [4], which resorts to the neural networks for building the mapping between the cross-domain data. Despite the observed visually satisfactory results, the 3D-R2N2 approach has the following limitations: 1) the brutal-force mappings between 3D shapes and natural images are unlikely to attend the fine-grained objects' poses and structures; 2) multiple images captured from different observation viewpoints are required as inputs to the network; 3) objects are always observed from natural images with cluttered backgrounds.

In this work, we aim to address above limitations by proposing a structure-aware 3D (SA3D) shape synthesis approach. Instead of requiring multiple image inputs, the proposed SA3D takes only a single image from a certain viewpoint as input, and synthesizes multiple "virtual" images from different viewpoints using the pre-trained deep generative models. In order to learn the deep generative models for cross-view image synthesis, we generate aligned object images from multiple viewpoints through projecting 3D shapes to 2D images with consistent projection parameters. When learning the mapping between multi-view images and corresponding 3D shapes, such aligned object poses across different views can facilitate the model's awareness towards the objects' geometric structures. Our main contributions are summarized as follows:

- We propose a SA3D shape synthesis approach that only requires a single-view image as the input. Meanwhile, the learned 2D-to-3D mapping is aware of the geometric structure of the objects in cross-view images when synthesizing the 3D objects.

- To our best knowledge, the SA3D approach is the pioneering work that attempts to employ synthesized 2D cross-view images for improving the quality of synthesized 3D shapes.

- Extensive quantitative and qualitative results demonstrate that the proposed approach can achieve the state-of-the-art performance on the ShapeNet datasets [2] with single-view image inputs.

2 Related Work

2.1 Reconstructing 3D shapes from images

How to reconstruct 3D shapes from a single or a handful of images has been a long standing problem. Some approaches [1, 7, 13] rely on matching image features from different-view images to reconstruct 3D shapes. For example, Huang *et al.* [13] estimated the viewpoints of large amounts of web images to match correspondences between the images and pre-existing models for 3D model reconstruction. However, these approaches are vulnerable to appearance changes of 3D shapes. As CNNs are getting popular in the image domain, a growing trend is to generate 3D shapes with CNN methods. Song *et al.,* [28] leveraged the coupled nature of the 3D shape completion task and the depth labeling task, and proposed an end-to-end 3D CNN that takes an depth image as input and generates complete 3D voxels with semantic labels. Wu *et al.* [32] generated volexilzed 3D models with the pre-trained network from depth images. More recently, Niu *et al.,* [24] proposed a convolutional-recursive autoencoder that generates cuboid structure of the parts of 3D models. Some other CNN-based learning approaches leverage large repositories of 3D CAD models. For example, Choy *et al.,* [4] proposed a unified 2D-to-3D learning approach for both single and multi-view based 3D object reconstruction with minimal supervision required. In addition to CNNs, probabilistic graphical models (*e.g.,* conditional random field [18]) are also employed for learning the 2D-to-3D mapping [20].

2.2 Deep generative models

Deep generative models have received extensive attentions in recent years, including generative adversarial networks (GANs) [9], variational autoencoders (VAE) [16] and their variants [22, 36]. Benefiting from both the GANs and volumetric convolutional networks, Wu *et al.,* [32] proposed a 3D generative adversarial network (3DGAN) that generates a 3D object from the probabilistic space. Based on the GAN framework, Gadela *et al.* [8] proposed PrGANs to train a projector, where the discriminator is trained to distinguish the projection images of real 3D models from those from generative models. Zhu *et al.* [35] proposed to introduce an enhancer that feeds the learned high-level feature images to the generator of the GAN to generate 3D models better. Inspired by prior arts that extended VAE for learning the mappings between cross-modality data [26], we propose to adapt VAE for synthesizing cross-view object images, and eventually enhance the quality of the single-view reconstructed 3D shapes with synthesized multi-view images.

3 Structure-Aware 3D shape Synthesis with Single-view Images

3.1 Single-View to Multi-View Object Image Synthesis

In order to learn a robust mapping between cross-view images, we utilize variational approach for latent representation learning, whereby aligned object poses across different views will be embedded into a unified latent space. In particular, we learn a deep generative model for each object category in an unsupervised fashion, where each deep generative model consists of an encoder network and a decoder network. The encoder network embeds all images into a compact vector representing the mean the data distribution in the latent feature space. At the meantime, a latent vector representing the variance of the training data is obtained through the encoder network as well. After that, the decoder network interprets these latent faces into cross-view pose object images with the same size of inputs of the encoder, so that the conversation between multi-view pose objects is effectively bridged.

The Kullback-Leibler (KL) divergence [16] is used to constrain the mean latent vector and the variance latent vector of the output of the encoder network is shown as follows:

$$L_{\text{kl}} = \frac{1}{2}\left[\, z_{\text{mean}}^2 + z_{\text{var}}^2 - \log\left(z_{\text{var}}^2 + eps\right)\right] - 1 \tag{1}$$

where, z_{mean} represents the mean latent vector, z_{var} represents the variance latent vector, eps is equal to $1e$-8. Furthermore, we adopt l_2 norm to represent the difference of pixel-wise of the generated images and the real images. Thus, we formulate the loss as:

$$L_{\text{recon}} = \frac{1}{2}\sum_{i}^{m} \parallel G(x)_i - y_i \parallel_2^2 \tag{2}$$

where, $G(x)$ represents generated target domain image, y represents real object target domain image. Eventually, we form our total loss function as:

$$L_{\text{total}} = L_{\text{kl}} + L_{\text{recon}} + \lambda L_{\text{reg}} \tag{3}$$

where, λ represents the regularization hyper-parameter, L_{reg} represents weight decay for all CNN parameters. We optimize our model by minimizing the L_{total}.

In order to enhance the synthesized image, we used U-Net [27] that a "fully convolutional network" [21]. The main idea in [27]: It will concatenate the output of the last deconvolution layer and the output of the corresponding encoder convolutional layers together and feed into the next deconvolution layer. In this way, a successive deconvolution layer can then get to assemble a more consummate output based this deep and shallow features. Figure 1 shows the pipeline of our model.

3.2 3D Shape Reconstruction from Synthesized Multi-View 2D Object Images

The 2D-to-3D generation network employed the same 3D recurrent module [10] as in [4], which is capable of utilizing both the current observation and previous observations, and can eventually improve the reconstruction quality as the observation view increased. More detailed network structure is as follows: Firstly, an encoder network employs standard 2D

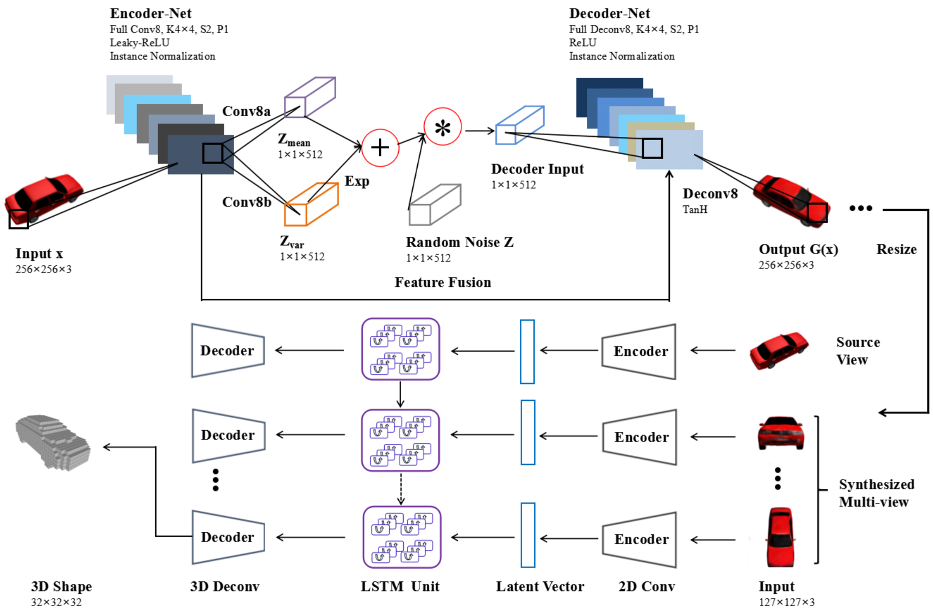


Figure 1: The pipeline of the proposed SA3D shape synthesis approach.

convolution layers, pooling layers and leaky rectified linear units to encode input images into lower dimensional latent vectors. Next, the 3D Convolution Gated Recurrent Units (GRUs) [3] is applied to the latent vectors for selectively updating cell memory status in GRUs. Finally, given the hidden states of the GRUs, the 3D Deconvolutional Neural Network can decode the input images and generates a 3D voxel reconstruction model.

The parameters 3D shape generation network are updated by minimizing the sum of voxel-wise cross-entropy. The loss function is expressed as follows:

$$L(\chi,y) = \sum_{i,j,k} y_{(i,j,k)} log(p_{(i,j,k)}) + (1-y_{(i,j,k)}) log(1-p_{(i,j,k)}) \tag{4}$$

where, χ represent the collection of input that omitted relevance. $p_{(i,j,k)}$ represent the final softmax layer output the occupancy probability $p_{(i,j,k)}$ of the voxel cell at (i,j,k). $y_{(i,j,k)} \in \{0,1\}$ represent the ground truth voxel occupancy.

3.3 Deep Network Modeling

For the cross-view image synthesis network, we adopt a three-step training procedure: Firstly, we train our network on all training sets instead of training the network on per-category training sets. Secondly, we add L_2 regularization loss to all parameters of the encoder network and the decoder network, and set the regularization hyper-parameter λ of 0.0005. Finally, dropout [12] regularization with a probability of 0.7 is added after the first three layers of the decoder network. In the encoder network, we use eight full convolutional layers structure (kernel size = 4, padding = 1, stride = 2) , and instance normalization [29] and Leaky-ReLU [33] activation functions are applied to all encoder network convolution layers. Leaky-ReLU slope is 0.2. In the decoder network, we use eight deconvolution layers structure (kernel size = 4, padding = 1, stride = 2), and instance normalization and ReLU [23] activation functions are applied to all decoder network deconvolution layers, excluding the output layer. Each layer of deconvolution output fused the output of the corresponding convolutional layer. Our models use the ADAM Optimizer [25] to set learning rate of 0.0002 and the momentum of 0.5. We set the training batch size to 64 and trained our model for 50,000 iterations. Table 1 shows our network architectures [36].

Net	Layer	Kernel Size	Stride	Activation Function	Output
Encoder-Net	Input	-	-	-	256 × 256 × 3
	Conv1 IN	4 × 4	2	Leaky-ReLU	128 × 128 × 64
	Conv2 IN	4 × 4	2	Leaky-ReLU	64 × 64 × 128
	Conv3 IN	4 × 4	2	Leaky-ReLU	32 × 32 × 256
	Conv4 IN	4 × 4	2	Leaky-ReLU	16 × 16 × 512
	Conv5 IN	4 × 4	2	Leaky-ReLU	8 × 8 × 512
	Conv6 IN	4 × 4	2	Leaky-ReLU	4 × 4 × 512
	Conv7 IN	4 × 4	2	Leaky-ReLU	2 × 2 × 512
	Conv8a	4 × 4	2	-	1 × 1 × 512
	Conv8b	4 × 4	2	-	1 × 1 × 512
Decoder-Net	Decoder Input	-	-	-	1 × 1 × 512
	Deconv1 IN	4 × 4	2	ReLU	2 × 2 × (512 × 2)
	Deconv2 IN	4 × 4	2	ReLU	4 × 4 × (512 × 2)
	Deconv3 IN	4 × 4	2	ReLU	8 × 8 × (512 × 2)
	Deconv4 IN	4 × 4	2	ReLU	16 × 16 × (512 × 2)
	Deconv5 IN	4 × 4	2	ReLU	32 × 32 × (256 × 2)
	Deconv6 IN	4 × 4	2	ReLU	64 × 64 × (128 × 2)
	Deconv7 IN	4 × 4	2	ReLU	128 × 128 × (64 × 2)
	Deconv8	4 × 4	2	TanH	256 × 256 × 3

Table 1: Architectures for the cross-view image generation network. Conv8a and Conv8b are shown in figure1. IN denotes instance normalization.

In training phase of 3D generation network, the input images are randomly cropped from the ShapeNet datasets [2], where the input size was set to 127 × 127 × 3, the output size was set to 32 × 32 × 32 that object voxelized reconstruction model. This network was trained up to 40,000 iterations with a batch size of 24 on a subset of the ShapeNet datasets (*include 50,000 objects and 13 major categories*). Based on the 40,000 iterations network parameters, we fine-tune 10,000 iterations of network parameters on the object single-view and synthesized multi-view datasets and use the ADAM Optimizer to set learning rate of 0.0001.

4 Experiments

In the experimental section, we will first introduce our experimental settings and then illustrate both qualitative and quantitative analysis results on 3D shape reconstruction. Our model is trained on a GPU server configured with two GTX TITAN X cards.

4.1 Dataset

ShapeNet: The ShapeNet datasets [2] is a popular 3D CAD datasets that has been widely used for 3D shape reconstruction, recognition and retrieval tasks. It consists of a large repository of 3D CAD models that are well-organized according to the WordNet hierarchy. In

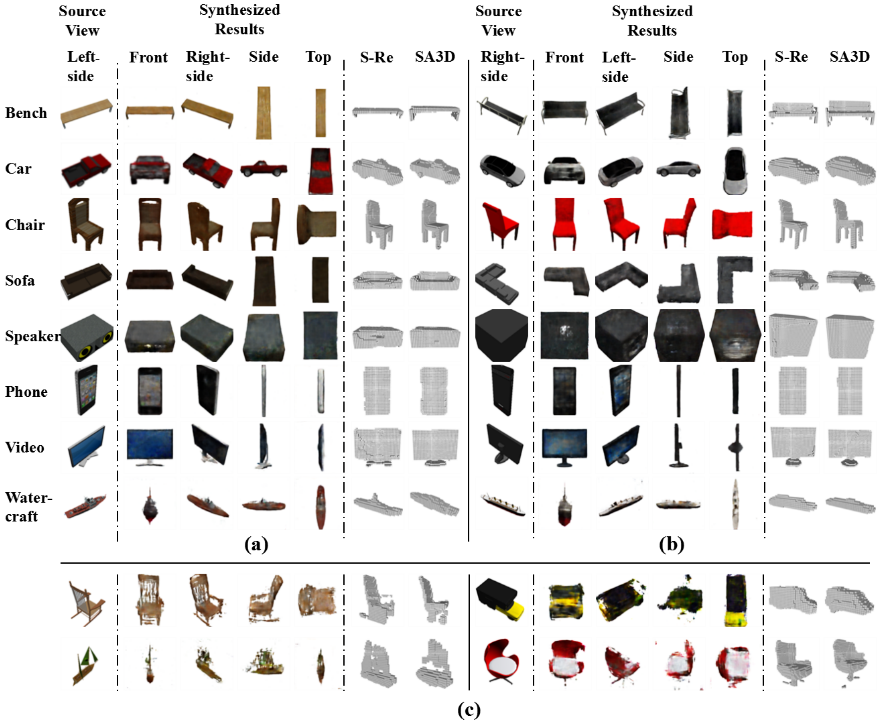


Figure 2: Reconstruction samples of single-view and synthesized multi-view. The left column of the dotted line is the real view of the object, and the right four columns are the synthesized views based on the object real view. (a) and (b) synthesized object view of four different perspectives based on object source view and reconstruction performance. (c) synthesized multi-view or reconstruction failure cases on testing sets. S-Re [4] indicates the reconstruction result of the single-view. SA3D indicates the reconstruction result of the synthesized multi-view.

our experiment, we selectively choose certain commonly seen object categories from the ShapeNet datasets. In order to build up the alignment of objects' poses across different views, we use a open-source tool kit ShapeNet-Viewer to generate 2D rendering images in batches for certain object categories (*include bench, car, chair, phone, sofa, speaker, video and watercraft*). More specifically, rendering images of each object are obtained from 5 selected views (*include front, leftside, rightside, side and top view* [1]). For all experiments, where the training sets contains 18,692 object samples and the testing sets contains 4,483 object samples. The training sets and the testing sets are randomly selected from the per-category at a ratio of approximately 4:1.

Evaluation criteria: Throughout all our experiments, the proposed SA3D approach is evaluated by comparing the Intersection-over-Union(IoU) of each object's single-view and synthesized multi-view reconstructed voxelization models with ground-truth voxelization models. The formula follows:

[1] Five different perspectives of the object. See Figure 2.

$$IoU = \sum_{i,j,k}[I(p_{(i,j,k)} > t)I(y_{(i,j,k)})] / \sum_{i,j,k}[I(I(p_{(i,j,k)} > t) + I(y_{(i,j,k)}))] \quad (5)$$

$I(\cdot)$ is the indicator function and t is the voxelization threshold in the 3D generation network, in order to let voxel *(i,j,k)* obey the Bernoulli distribution. Other variables are introduced in section 3.2. The higher IoU values indicate that voxel reconstruction performance better. Next, we will compare the single-view and synthesized multi-view IoU reconstruction performance.

4.2 Evaluation

We report both qualitative and quantitative evaluations of single-view and synthesized multi-view reconstruction performance on the all testing sets.

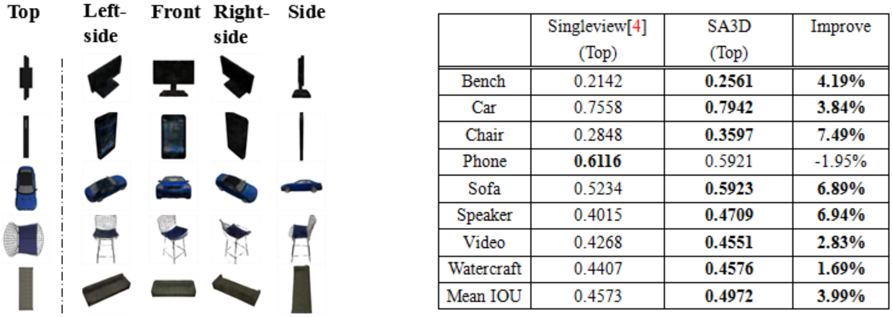
Experiment setup: We respectively select the leftside view and the rightside view as the source view since these two views are relatively more informative than the other three views. We fine-tune 10,000 iterations 3D generation network on all single-view and synthesized multi-view training sets. At the same time, we set the same hyper-parameter t (voxelization threshold) = 0.4 in fine-tune process, t is defined in formula 5. Finally, we respectively use 3D generation network get single-view 3D voxel reconstruction model and synthesized multi-view 3D voxel reconstruction model on all testing sets. Note that in all experiments, our method does not require any category labels.

Qualitative results: Figure 2 illustrates examples of the authentic source view images, the synthesized cross-view images and the corresponding reconstructed 3D shapes in the testing sets, which strictly does not overlap with the training sets. More specifically, Figure 2 (a) and (b) show some example reconstructions from the object single-view and structure-aware 3D shape synthesized results. We can draw the following analysis and conclusion. The cross-view synthesize network can generated multi-view of objects, and simultaneously the decoder layer fusion the object's features, the synthesized views has better performance in details such as object outline, color, and length, etc. Since the synthesized multi-view can better represent the geometric structure of the object as well as containing more details, it has better reconstruction performance than single-view. We also illustrate some failure cases in Figure 2 (c).

	Singleview[■] (Rightside)	SA3D (Rightside)	Improve	Singleview[■] (Leftside)	SA3D (Leftside)	Improve	Mean Improve
Bench	**0.4392**	0.4272	-1.20%	0.3903	**0.4325**	**4.22%**	**1.51%**
Car	0.7986	**0.7993**	**0.07%**	0.7694	**0.7968**	**2.74%**	**1.40%**
Chair	**0.4947**	0.4853	-0.94%	0.4755	**0.4945**	**1.90%**	**0.48%**
Phone	0.6798	**0.7692**	**8.94%**	0.6869	**0.7813**	**9.44%**	**9.19%**
Sofa	**0.6425**	0.6364	-0.61%	0.6315	**0.6543**	**2.28%**	**0.83%**
Speaker	0.6653	**0.6915**	**2.62%**	0.6792	**0.7024**	**2.32%**	**2.47%**
Video	0.5494	**0.5778**	**2.84%**	0.5219	**0.5750**	**5.13%**	**3.98%**
Watercraft	**0.5322**	0.4832	-4.90%	**0.5375**	0.4931	-4.44%	-4.67%
Mean IoU	0.6002	**0.6087**	**0.85%**	0.5865	**0.6162**	**2.97%**	**1.91%**

Table 2: Compare per-category reconstruction results for single-view and synthesized multi-view on all testing sets. These values are Mean IoU.

Quantitative results: As described above, we select two views (*leftside and rightside*) out of the total 5 views as the source views. All parameters are set consistently throughout the experiment. Table 2 shows per-category single-view and synthesized multi-view reconstruction results. Based on object rightside view, our method improve four categories, we improve the IoU by **0.85**% on average. Based on object leftside view, our method improve seven categories, we improve the IoU by **2.97**% on average. Our method worse on watercraft categories, because it has lots of details and higher shape variation. In all experiments, we improve the IoU by **1.91**% on average.



	Singleview[4] (Top)	SA3D (Top)	Improve
Bench	0.2142	**0.2561**	**4.19%**
Car	0.7558	**0.7942**	**3.84%**
Chair	0.2848	**0.3597**	**7.49%**
Phone	**0.6116**	0.5921	-1.95%
Sofa	0.5234	**0.5923**	**6.89%**
Speaker	0.4015	**0.4709**	**6.94%**
Video	0.4268	**0.4551**	**2.83%**
Watercraft	0.4407	**0.4576**	**1.69%**
Mean IOU	0.4573	**0.4972**	**3.99%**

(a) Samples

(b) Per-category IoU

Figure 3: (a). The synthesized multi-view samples based on the object top view. (b). Per-category reconstruction IoU comparison.

Figure3 (a) shows partial synthesized multi-view samples based on object top view. Figure3 (b) shows the results of single-view and synthesized multi-view based on top view that reflect object local information. Figure 3 shows that our method can improve the performance of 3D reconstruction when we only observe the local information of the object. In this experiments, we train 20,000 iterations of 3D generation networks in object top view and synthesized multi-view training sets. Our method enhances seven categories reconstruction performance, and improve the IoU by **3.99**% on average compared to single-view. Table 3 shows based on object leftside view comparison. SA3D-3 represent three views (leftside, front and rightside) generated based on the leftside view. SA3D-5 represent five views (leftside, front, rightside, side and top) generated based on the object leftside view. We found that the reconstruction quality improve for seven categories as the number of views. SA3D-3 IoU have an average improved of **2.32**% compared to single-view IoU. SA3D-5 IoU have an average improved of **2.97**% compared to single-view IoU. Bench, Phone and Video had the highest reconstruction improved since these objects have fewer shape changes.

	Singleview[■] (Leftside)	SA3D-3 (Leftside)	Improve	SA3D-5 (Leftside)	Improve
Bench	0.3903	**0.4478**	**5.75**%	**0.4325**	4.22%
Car	0.7694	**0.8014**	**3.20**%	0.7968	2.74%
Chair	0.4755	0.4782	0.27 %	**0.4945**	**1.90**%
Phone	0.6869	0.7525	6.56 %	**0.7813**	**9.44**%
Sofa	0.6315	**0.6666**	**3.51**%	0.6543	2.28%
Speaker	0.6792	0.6986	1.94 %	**0.7024**	**2.32**%
Video	0.5219	0.5388	1.69 %	**0.5750**	**5.13**%
Watercraft	**0.5375**	0.4941	-4.32%	0.4931	-4.44%
Mean IoU	0.5865	0.6097	2.32%	**0.6162**	**2.97**%

Table 3: Compare per-category reconstruction results for Singleview, SA3D-3 and SA3D-5 on all testing sets.

5 Conclusion

In this paper, we addressed some common restrictions of image-based 3D shape reconstruction approaches, and proposed a structure-aware 3D shape synthesis method that requires only a single-view image input. To our best knowledge, the proposed method is the first attempt that employs synthesized multi-view images for enhancing the quality of single-view reconstructed 3D shapes. In order to guide the learned model to be aware of the geometric structure of objects, we generated aligned object poses across multi-view images through projecting 3D shapes to 2D images with consistent projection parameters. We provided extensive quantitative and qualitative results on selected categories of the ShapeNet datasets, where the proposed SA3D approach can lead to an average of **2.32%** performance improvement over the single-view based baseline with 2 synthesized views, and an average of **2.97%** performance improvement with 4 synthesized views.

Acknowledgment

The work of J. Tang was supported by the National Natural Science Foundation of China under Grant 61772032.

References

[1] Caroline Baillard, Cordelia Schmid, Andrew Zisserman, and Andrew Fitzgibbon. Automatic line matching and 3d reconstruction of buildings from multiple views. In *ISPRS Conference on Automatic Extraction of GIS Objects from Digital Imagery*, volume 32, pages 69–80, 1999.

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

[3] Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *Computer Science*, 2014.

[4] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European Conference on Computer Vision*, pages 628–644. Springer, 2016.

[5] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015.

[6] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[7] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *European Conference on Computer Vision*, pages 834–849. Springer, 2014.

[8] Matheus Gadelha, Subhransu Maji, and Rui Wang. 3d shape induction from 2d views of multiple objects. *CoRR*, abs/1612.05872, 2016.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[10] Alex Graves. *Long Short-Term Memory*. Springer Berlin Heidelberg, 2012.

[11] Ankur Handa, Thomas Whelan, John McDonald, and Andrew J Davison. A benchmark for rgb-d visual odometry, 3d reconstruction and slam. In *Robotics and automation (ICRA), 2014 IEEE international conference on*, pages 1524–1531. IEEE, 2014.

[12] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *Computer Science*, 3(4):pÃągs. 212–223, 2012.

[13] Qixing Huang, Hai Wang, and Vladlen Koltun. Single-view reconstruction via joint analysis of image and shape collections. *ACM Trans. Graph.*, 34(4):87:1–87:10, 2015. doi: 10.1145/2766890.

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

[15] Hanme Kim, Stefan Leutenegger, and Andrew J Davison. Real-time 3d reconstruction and 6-dof tracking with an event camera. In *European Conference on Computer Vision*, pages 349–364. Springer, 2016.

[16] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[18] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. pages 282–289, 2001.

[19] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 239–248. IEEE, 2016.

[20] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5162–5170, 2015.

[21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[22] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. *arXiv preprint arXiv:1701.04722*, 2017.

[23] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *International Conference on International Conference on Machine Learning*, pages 807–814, 2010.

[24] Chengjie Niu, Jun Li, and Kai Xu. Im2struct: Recovering 3d shape structure from a single rgb image. *arXiv preprint arXiv:1804.05469*, 2018.

[25] Frank A Nothaft, Matt Massie, Timothy Danford, Zhao Zhang, Uri Laserson, Carl Yeksigian, Jey Kottalam, Arun Ahuja, Jeff Hammerbacher, Michael Linderman, Michael Franklin, Anthony D. Joseph, and David A. Patterson. Rethinking data-intensive science using scalable analytics systems. In *Proceedings of the 2015 International Conference on Management of Data (SIGMOD '15)*. ACM, 2015.

[26] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions. In *Advances in neural information processing systems*, pages 2352–2360, 2016.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. URL http://arxiv.org/abs/1505.04597.

[28] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 190–198. IEEE, 2017.

[29] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. 2016.

[30] Ting-Chun Wang, Alexei A Efros, and Ravi Ramamoorthi. Occlusion-aware depth estimation using light-field cameras. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 3487–3495. IEEE, 2015.

[31] Jiajun Wu, Tianfan Xue, Joseph J Lim, Yuandong Tian, Joshua B Tenenbaum, Antonio Torralba, and William T Freeman. Single image 3d interpreter network. In *European Conference on Computer Vision*, pages 365–382. Springer, 2016.

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

[33] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *CoRR*, abs/1505.00853, 2015. URL http://arxiv.org/abs/1505.00853.

[34] Qian-Yi Zhou and Vladlen Koltun. Color map optimization for 3d reconstruction with consumer depth cameras. *ACM Transactions on Graphics (TOG)*, 33(4):155, 2014.

[35] Jing Zhu, Jin Xie, and Yi Fang. Learning adversarial 3d model generation with 2d image enhancer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, 2018.

[36] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.