Pacific Graphics (2018)
H. Fu, A. Ghosh, and J. Kopf (Editors)

Short Paper

3D VAE-Attention Network: A Parallel System for Single-view 3D Reconstruction

Fei Hu, Xinyan Yang , Wei Zhong, Long Ye, and Qin Zhang

Key Laboratory of Media Audio & Video , Communication University of China

Abstract

3D object reconstruction from single view image is a challenge task. Due to the fact that the information contained in one isolated image is not sufficient for reasonable 3D shape reconstruction, the existing results on single-view 3D reconstruction always lack marginal voxels. To tackle this problem, we propose a parallel system named 3D VAE-attention network (3VAN) for single view 3D reconstruction. Distinct from the common encoder-decoder structure, the proposed network consists of two parallel branches, 3D-VAE and Attention Network. 3D-VAE completes the general shape reconstruction by an extension of standard VAE model, and Attention Network supplements the missing details by a 3D reconstruction attention network. In the experiments, we verify the feasibility of our 3VAN on the ShapeNet and PASCAL 3D+ datasets. By comparing with the state-of-art methods, the proposed 3VAN can produce more precise 3D object models in terms of both qualitative and quantitative evaluation.

CCS Concepts

•***Computing methodologies*** → *Reconstruction; Volumetric models;*

1. Introduction

3D reconstruction is an integral problem in geometric computing and modeling. There already have been considerable researches in 3D reconstruction based on images.In many cases, we need to recover 3D shape from single-view image. The single-view 3D reconstruction is an ill-posed problem due to the lack of disparity information, so that the traditional methods require additional prior knowledge.

Owing to the remarkable achievement of learning methods and the establishment of various 3D object databases, learning methods have been gradually introduced into 3D reconstruction tasks. [GFRG16] [TDB16] Although the above methods can be applied to perform the 3D reconstruction tasks, they have not given the subjective and objective evaluations. Choy et al. [CXG*16] proposed an overall framework called 3D-R2N2 for single-view and multi-view reconstruction based on LSTM which means it had high computation complexity. Fan H et al. [FSG17] proposed a point set generation network for 3D object reconstruction from one single image.

Especially when we reproduce the experiments of [CXG*16] [FSG17], we find that the output models often lack detailed information on single-view reconstruction task. We attribute the problem to the ill-conditioned nature of the single-view reconstruction task and the one-sidedness of the current networks. And parts of missing information can be found in the corresponding image. So we propose to address this information missing by introducing attention mechanism to the task.

Our main contribution is to build an end-to-end parallel system 3D VAE-attention network (3VAN) for single view 3D reconstruction task. Our proposed 3VAN consists of two branches. The first branch learns to generate 3D rough shape of an object. We feed the corresponding image into modified 3D variational autoencoder reconstruction architecture to get the general volumetric occupancy. The other one integrates the details of the 3D object by attention mechanism. It learns to endow higher weights to the features of missing details in the image. Consequently, in the Attention Network, we can obtain volumetric occupancy which represents the details of object. Finally, we put the volumetric occupancy of these two branches together to get the full 3D shape object model. Our architecture generates 3D object models which contain more vivid details and makes qualitative and quantitative improvements on ShapeNet dataset [YSS*17] compared with [CXG*16] and [FSG17]. The main contributions of this paper are summarized as follows:

- We propose a parallel system instead of the common encoder-decoder architecture.
- We introduce attention mechanism to the 3D reconstruction task.
- We propose an extension of the standard VAE generator framework in our contour reconstruction branch.

DOI: 10.2312/pg.20181279

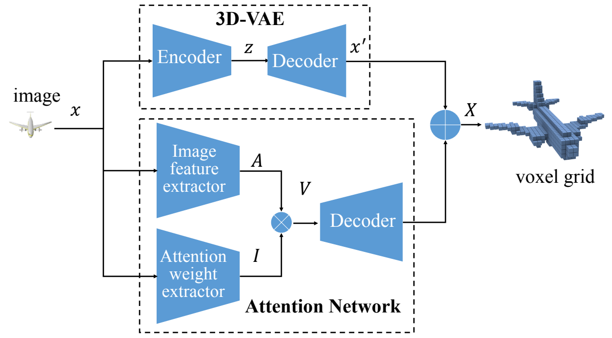

Figure 1: *The general framework of 3VAN*

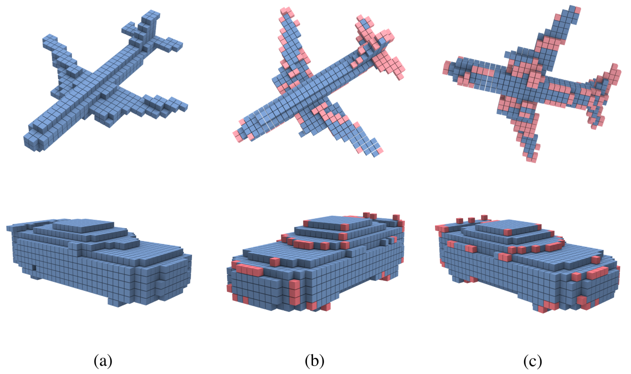

Figure 2: *A sample of the two components reconstruction visualization. The first row is airplane and the second row is car. (a) is the ground truth model in the ShapeNet dataset. (b) and (c) are the predictions of 3VAN in different projection, in which the blue voxel girds are produced by the 3D-VAE and the pink ones are made up by Attention Network.*

- We propose a reasonable combined loss function for our network.
- On the task of single-view reconstruction, our framework outperforms state of the art.

2. 3D VAE-Attention Network

In this section, we propose an effective 3D VAE-Attention Network to reconstruct authentic 3D object model from one single view image,. It decomposes the 2D-to-3D reconstruction task into two branches (as shown in Figure 1).

- **3D-VAE**: This branch extends the standard VAE for adapting to 3D reconstruction tasks. It reconstructs rough 3D object shape conditioned on a single image which is sampled from an arbitrary view, yielding an uncompleted volumetric occupancy.
- **Attention Network**: By inputting the same single view image, it makes up the defects in uncompleted volumetric occupancy and reconstructing detailed voxel occupancy to integrate the 3D shape.

Eventually, we combine these two aforementioned outputs together to get the completed 3D shape of object model. The combination is visualized and shown in Figure 2.

2.1. Variational Autoencoder

Variational Autoencoder is composed of two networks that one encodes the input data x to the latent vector z and the other decodes the latent vector back to the data space for target generation $\widetilde{x}$. The process of these encode-decode system can be simplified as the following equation:

$$z = encoder(x) = q(z \mid x), \tag{1}$$

$$\widetilde{x} = decoder(z) = p(\widetilde{x} \mid z), \tag{2}$$

where x and $\widetilde{x}$ are both observed distribution during training and z is learned to represent the mapping relation between them.

For our 3D-VAE framework, a 2D image x sampled from an arbitrary view is fed into the 2D-image-encoder to produce low dimensional feature vector z. The 3D-voxel-decoder expanses the image feature to generate the corresponding 3D volumetric occupancy $\widetilde{x}$. We pre-train the 3D-rough-shape generation branch with random-sampling which is a VAE structure. During the training process of upper-branch, we initialize the encoder with the pre-trained parameters and remove the random-sampling for better performance of the whole architecture.

2.2. Attention network

Inspired by the fact that the volumetric occupancy generated from 3D-VAE network lacks marginal voxels often, we add an attention branch to the 3D reconstruction for completing the shape of 3D models. We design a fully convolutional Attention Network as shown in the lower half branch of Figure 1 to establish the correspondence between missing details in volumetric occupancy and the local feature of image.

Convolutional network can extract a set of feature vectors from the 2D image. The extractor produces an m-dimensional vector $A = \{a_1, a_2, \cdots, a_m\}$, $a_m \in \mathcal{R}$, each element of which represent a local region of the image.

As the 3D-VAE can produce a rough shape of volumetric occupancy, we can get the residual voxel occupancy which fail to reconstruct. In the convolutional Attention Network, the backpropagation algorithm will project the residual voxels back to the attention weight eigenvector. The mapping relation is recorded as I and can be obtained from an interlayer. $I = \{i_1, i_2, \cdots, i_n\}$, $i_n \in \mathcal{R}$ is a n-dimensional vector and each element symbolizes the pertinence between these local regions of image and the residual voxels. So that it can be regarded as importance weighted eigenvector.

As shown in the lower half branch of Figure 1, we feed an 2D image x into the Attention Network. From two sub-branches (i.e. regarded image feature extraction as weight matrixW_f and importance weighted extraction as weight matrix W_ω) we can get a attention vector V when $m = n$. V contains the information that the different regions of image contribute to shape completion variously. After feeding into the decoder (W_ω), V is expanded to the residual

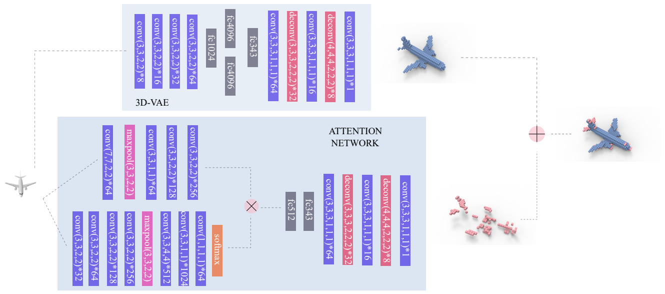

Figure 3: *The implementation architecture of 3VAN*

voxels. This process can be formulated in the following formulation and X is the distribution of full shape 3D model,

$$V = W_f x \odot (W_\omega x)^T = A \odot I, \tag{3}$$

$$X - \widetilde{x} = W_d \cdot V. \tag{4}$$

Attention Network is voxel complemental branch of 3VAN and shares the same input image with 3D-VAE branch. It establishes the mapping relation between the local feature of image and the missing marginal voxels, and utilize the relationship to endow higher weight to more crucial region of image. Through the same decoding architecture as 3D-VAE branch, Attention Network produces detailed volumetric occupancy utimately.

2.3. The proposed network architecture

The network architecture pipeline is shown in Figure 3. We feed a single image which is sampled from arbitrary view into the two branches of 3VAN. In the upper half branch of Figure 3, 3D-VAE network first uses several 2D convolutional layers which encode the 127 × 127 image into an eigenvector. Then, the decoder which is composed by 3D convolutional and deconvolutional layers extends the 343 dimensional eigenvector into $32 \times 32 \times 32$ voxel occupancy of 3D rough shape. In the lower half branch of Figure 3, the Attention Network endows high weights to unreconstructed regions by attention mechanism which is a cocurrent convolutional network and uses the same decoder architecture to get the $32 \times 32 \times 32$ voxel occupancy of 3D detailed shape. In attention mechanism module, the upper sub-branch extracts the 2D image feature and the lower sub-branch learns the attention weight eigenvector. For clear illustrate, here we take a sample of airplane category as an example. The results of these two components of attention are shown in Figure 2. The blue voxels describe a rough 3D shape and the pink ones add the details. We sum the results of these two components up for getting more precise 3D shape of object model.

2.4. Loss function

Loss function is a critical factor for the convergence of neural network. For our reconstruction task, we add a quadratic component to a common loss function weight Sigmoid Cross Entropy with Logits ($\omega - SCE$), and the loss function could be expressed as:

$$Loss = -\omega\, t \log(\sigma(o)) - (1-t)\log(1-\sigma(o)) + \lambda(t - \sigma(o))^2, \tag{5}$$

where $\sigma(x) = \frac{1}{1-e^{-x}}$, $o = w_{ij}x_{ij} + b_{ij}$, t is the representation of occupancy which is either 0 or 1, o is the output of the network , ω and λ are hyper parameters. In our paper, $\omega = 20$, $\lambda = 1$.

3. Experiments

In this section, we firstly introduce the dataset and training details, and verify the feasibility of 3VAN in quantitative (summarized in the Table 1 and Table 2) and qualitative (shown in Figure 2).

3.1. Datesets and implementation details

Dataset The ShapeNet dataset is a richly-annotated, large-scale dataset of 3D shapes. It is collected by Princeton, Stanford and TTIC. We use a subset of the ShapeNet dataset which contains about 40,000 3D models over 13 common categories.

We evaluate the generalization ability of 3VAN on the PASCAL 3D+ dataset [XMS14]. We convert the CAD model to voxel format. This dataset contains 12,093 3D models over 12 common categories. We split the dataset into training and testing sets, with $1/2$ for training and the remaining $1/2$ for testing.

Implementation details We use the ADAM [KB14] solver for stochastic optimization in all the experiments. During the training time, the learning rate is 10^{-4} (whole network) or 10^{-3} (pretrain) for the neural networks. The representation of accuracy is IOU. We pretrain the 3D-VAE, and use these pretrained parameters to initialize the 3D-VAE before the whole network 3VAN is trained.

3.2. Quantitative and qualitative results

To validate our network structure, We compare our results with the state-of-the-art deep learning 3D reconstruction methods [CXG*16] [FSG17] on ShapeNet dataset. The IoU accuracy of 13 categories for our 3VAN and compared methods are reported in Table 1.

3D-R2N2 reconstructs 3D model from single or multi-view images, and in the single view reconstruction setting we achieved better performance in most categories. Particularly, in 8 out of 13 categories, our results are even better than 3D-R2N2 reconstructs for 5 views. Although Point-Net achieves higher IoU accuracy in some categories, the generated 3D model is represented by fixed amount of 2^{10} points, distinct from 3VAN reconstruct the 3D model with 2^{15} dimensional occupancy problem, which means the comparsion of evaluation is unfair and higher IoU accuracy in Point-Net is not positive correlation to model look more immersive. We train the network by all categories without any semantic labels. It causes the cross-impact effect, which leads to the performance decline of some categories. But the total performance of all categories gets better. The category-wise IoU of the 3D-VAE is 0.627, and that of 3VAN reaches to 0.640 indicate the attention mechanism works. In Table 2, 3VAN also outperforms the state-of-art methods on the PASCAL 3D+ dataset which verifies the generalization of our method.

	3D-R2N2			Point Net	3D VAE	3VAN
Viewpoint	1	3	5	1	1	1
Plane	0.513	0.549	0.561	0.601	0.565	**0.594**
Bench	0.421	0.502	0.527	0.55	0.499	**0.74**
Cabinet	0.716	0.763	**0.772**	0.771	0.755	0.597
Car	0.798	0.829	**0.836**	0.831	0.828	0.709
Chair	0.466	0.533	0.55	0.544	**0.605**	0.562
Monitor	0.468	0.545	0.565	0.552	**0.591**	0.59
Lamp	0.381	0.415	0.421	0.462	0.574	**0.771**
Speaker	0.662	0.708	0.717	**0.737**	0.715	0.566
Firearm	0.544	0.593	0.6	**0.604**	0.508	0.547
Couch	0.628	0.69	0.706	**0.708**	0.693	0.588
Table	0.513	0.564	0.58	0.606	0.598	**0.716**
Cellphone	0.661	0.732	0.754	0.749	0.697	**0.83**
Watercraft	0.513	0.596	0.61	**0.611**	0.535	0.513
Mean	0.56	0.617	0.631	0.64	0.627	**0.64**

Table 1: *3D reconstruction IoU on the ShapeNet dataset test*

method	IoU
Kar. [KTCM15]	0.318
3D-R2N2	0.517
3VAN	**0.6**

Table 2: *3D reconstruction IoU on the PASCAL 3D+ dataset test*

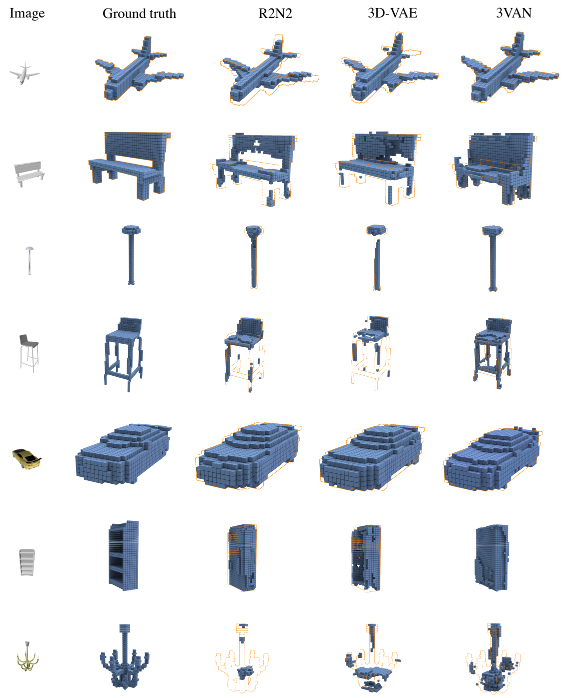

Figure 4: *Visualization of volumetric occupancy in distinct methods*

The voxel grid visualization of our experimental results is shown in Figure 4, we compare our reconstruction results with 3D-R2N2 and 3D-VAE for the qualitative analysis. The first five reconstruction samples show that 3VAN makes up the lacking details, especially in the thin structure such as the leg of a chair in the 4th row, so that we can get more precise results. And the last two rows show the limitation of the state-of-art methods in reconstructing hollow structure object which need to be further researched.

4. Conclusion

In this paper, we design a 3D reconstruction network 3VAN. The proposed method decomposes the prediction into two branches. The first one is 3D-VAE which produces rough 3D shape by an extension of standard VAE. The other one is Attention Network which establishes the correspondence between missing details in volumetric occupancy and regions in image to add the details for completing 3D model shape. By comparing with state-of-art methods and analyzing the structure effectiveness, 3VAN is verified to produce more precise 3D object models in qualitatively and quantitatively.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant Nos. 61631016 and 61371191, and the Project of State Administration of Press, Publication, Radio, Film and Television under Grant No. 2015-53 and the Fundamental Research Funds for the Central Universities No.2018XNG1825.

References

[CXG*16] Choy C. B., Xu D., Gwak J., Chen K., Savarese S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European Conference on Computer Vision* (2016), Springer, pp. 628–644. 1, 3

[FSG17] Fan H., Su H., Guibas L.: A point set generation network for 3d object reconstruction from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), vol. 38. 1, 3

[GFRG16] Girdhar R., Fouhey D. F., Rodriguez M., Gupta A.: Learning a predictable and generative vector representation for objects. In *European Conference on Computer Vision* (2016), Springer, pp. 484–499. 1

[KB14] Kingma D. P., Ba J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014). 3

[KTCM15] Kar A., Tulsiani S., Carreira J., Malik J.: Category-specific object reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 1966–1974. 4

[TDB16] Tatarchenko M., Dosovitskiy A., Brox T.: Multi-view 3d models from single images with a convolutional network. In *European Conference on Computer Vision* (2016), Springer, pp. 322–337. 1

[XMS14] Xiang Y., Mottaghi R., Savarese S.: Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision (WACV)* (2014). 3

[YSS*17] Yi L., Su H., Shao L., Savva M., Huang H., Zhou Y., Graham B., Engelcke M., Klokov R., Lempitsky V., et al.: Large-scale 3d shape reconstruction and segmentation from shapenet core55. *arXiv preprint arXiv:1710.06104* (2017). 1