Learning Single-view Object Reconstruction with Scaling Volume-View Supervision

Zishu Gao[1,2], Guodong Yang[1], En Li[1] and Zize Liang[1]
[1] State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100190, China
Email: {gaozishu2016, guodong.yang, en.li, zize.liang} @ia.ac.cn

***Abstract*—Producing a 3D voxel from a single view by deep learning-based methods has garnered increasing attention. Several state-of-the-art works introduce the recurrent neural network(RNN) to fuse features and generate full volumetric occupancy. However, the inputs are unable to be fully exploited to improve the reconstruction due to long-term memory loss. And most of the works have considered using 3D supervision for the whole optimization to recover the full volume, but lack detailed silhouette supervision to refine the reconstruction process. To address these issues, an end-to-end object reconstruction network with scaling volume-view supervision is proposed. We introduce an auto-encoder 3D volume predicting network that takes a single arbitrary image as input and outputs a voxel occupancy grid. And a scaling volume-view supervision module, which uses up-sampling to zoom errors and increase penalties, is leveraged to improve both the global and local optimization. Extensive experimental analysis on ShapeNet dataset shows that our network has superior performance when the scaling volume-view supervision is involved and the deep residual module boosts the reconstruction performance and speeds up the optimization effectively.**

I. Introduction

Object reconstruction is an essential task in many applications such as virtual reality, object manipulation and augmented reality. Recovering the full 3D volume from a single view is a very challenging task. Many attempts have been made using shape-from-X methods, such as shape-from-silhouette [1] [2], shape-from-shading [3] [4] and shape-from-texture [5]. The deformation of the texture easily affects the projection result and makes the reconstruction effect worse in [5]. The limitation these methods suffer from is that they require strong assumptions and expertise.

With the success of deep learning-based techniques, there are a number of researchers concern deep learning-based reconstruction methods [6] [7] [8]. First, many works use a recurrent neural network to refine the features extracted by the encoder from multiple views and achieve 3-dimensional reconstruction. Kar et al. introduce recurrent grid fusion to retain previous observations and refine the 3D output using 3D convolutional Gated Recurrent Unit (GRU) [9]. Based on the same philosophy, 3D-R2N2 leverages Long Short-Term Memory (LSTM) and GRU to fuse the features extracted from a sequence of images [10]. Some researchers use 3D-R2N2 as a baseline and integrate valuable features based on RNN methods [11] [12]. However, these methods need to solve the time-consuming problem. In addition, because of long-term memory loss, the inputs are unable to be full exploited to improvement the reconstruction.

In addition, Approaches based on generative adversarial networks (GANs) and variational autoencoders (VAEs) are proposed by a number of authors. MP-GAN adopts multiple discriminators that encode the distribution of 2D projections of the 3D shapes seen from different views [13]. 3D-RecGAN leverages the conditional adversarial network to recover the occluded regions by taking multiple images as input [14]. Wu et al. combine the GAN with the VAE to sample objects without a reference image or CAD models. However, they are limited by class labels for prediction [15]. Gal et al. propose a conditional adversarial loss and geometric adversarial loss to make a prediction using point cloud representations [16]. But these methods are limited to time-consuming and the instability of adversarial generation.

There are several representations of reconstruction used in many works. O-CNN [17] and OGN [18] introduce deep convolutional neural networks and use octree to present the reconstruction. However, octree presentation is so complex that will consume many computing resources. [19] introduces an approximate gradient for rasterization that enables the neural networks to produce 3D mesh from a single image. Pixel2Mesh [20] and Image2Mesh [21] produces a 3D shape in mesh from a single color image, but it is hard to transform 3D unstructured meshes into regular shapes. In addition, DensePCR introduces a deep pyramidal network for point cloud reconstruction [22]. 3D-LMNet proposes a latent embedding matching approach for 3D reconstruction in the point cloud presentation [23]. PSGN also generates point clouds taking a single view as inputs [24]. But due to the limited connections between points, the point cloud presentation is inaccurate overall. DeepShapeSketch [25] and 3D-INN [26] produce a freehand line drawing sketches from a 3D object under a given viewpoint, but this is challenging and limited to inaccuracy.

The works mentioned above suffer from the inaccuracy due to the long-term memory loss or typical output representations, and the instability of the adversarial framework. To avoid these problems, in this paper, we implement a novel auto-encoder network with scaling volume-view supervision, which outputs volumetric representation for single-view reconstruction. The

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

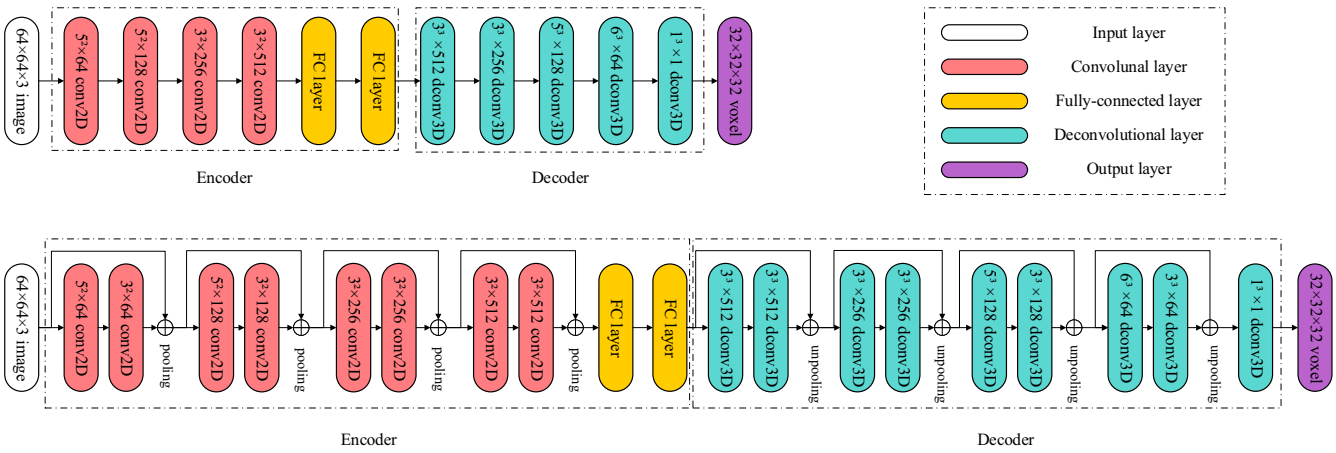

Fig. 1. The network structures for reconstruction from a single view. Top: a shallow network named VVSNet. Bottom: a network using deep residual module called VVSNet-R. The residual module helps to assist the feature extraction and improve the optimization process.

network contains three modules: encoder, decoder, scaling volume-view supervision. The scaling volume-view supervision module lets the network adopt both the volume and view optimization in the training process. When at the test time, the network can generate a full volumetric occupancy from a single image without the scaling volume-view supervision module. To achieve a good balance between accuracy and model size, we propose two versions of the networks.

There are two main contributions in this research of reconstruction framework,

- A 2d-to-3d framework for single-view reconstruction is established, and the deep residual framework shows a good performance in predicting 3D geometry.
- We present a scaling volume-view supervision that improves the accuracy of results by zooming errors and increasing penalties. At the same time, combining both the view and volume supervision helps to optimize the framework from both the global and local perspectives and make improvements on accuracy.

The remainder of the paper is structured as follows: In Section 2, the pipeline of our reconstruction framework is introduced firstly. Then the optimization strategy is designed in Section 3. The experimental evaluation and discussions are proposed in Section 4, and the conclusion and future works are given in Section 5.

II. Network Structure

In this section, we implement a unified end-to-end framework with the scaling volume-view supervision for single-view reconstruction. The 3D output of an object is represented by a voxel occupancy grid. Fig.1 shows the detailed network architectures in two versions of VVSNet and VVSNet-R. The former framework has fewer parameters, while the latter involves more parameters due to residual module, which can make more accurate 3D predictions and accelerate the optimization process.

A. Encoder

The encoder is utilized to extract a lot of informative features from a single image. As shown in Fig.1, we design two different encoders: a shallow VVSNet and a residual variation of it VVSNet-R. There are four 2D convolutional layers and two fully connected layers in VVSNet. The kernel size of the first two layers is 5^2, and the kernel size of the other two is 3^2. These convolution layers have the stride is 2. Each convolutional layer is followed by a batch normalization (BN) and a LeakyReLU activation. LeakyReLU activation is utilized to guarantee that neurons will not die when the input is less than 0, and BN is adopted to accelerate the convergence process. The number of output channels of these convolutional layers starts with 64 and double for the subsequent layers and ends up with 512. The output is passed to two fully connected layers and compressed into a 512-dimensional feature vector.

In VVSNet-R, we implement a deep residual version. The residual module plays an important role in assisting feature extraction and improving the optimization process. As shown in Fig. 1, we add 4 convolutional layers whose kernel size is 3^2 on the basis of the standard network. There is a 1×1 convolution as a residual connection between two convolution layers. The stride of these convolution layers is 1 and there is a pooling layer between two residual modules. The output channels of each residual module are 64, 128 and 512. After two fully connected layers, the feature vector is also of sizes 512.

B. Decoder

The decode takes feature maps as inputs and generates full 3D volumetric occupancy. In the standard framework VVSNet, there are five 3D transposed convolution layers. The kernel sizes of the first four transposed convolution layers are 3^3, 3^3, 5^3 and 6^3 with a stride of 2, and the output channels of these four layers are 512, 256,128 and 64, respectively. A BN

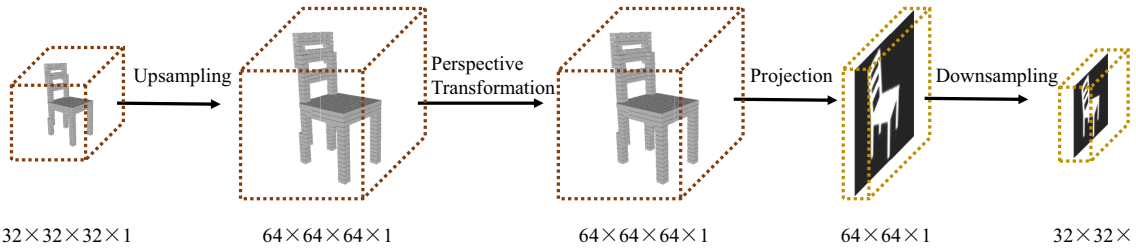

Fig. 2. An overview of the scaling volume-view supervision module. The module is composed of an up-sampling layer, perspective transformation, projection operation, and a down-sampling layer. It can zoom errors and increase the penalties using up-sampling layer for better performance.

layer and a ReLU activation are behind each of the first four convolutional layers. The last transposed convolutional layer is a kernel size of 1^3 in order to change the output channel to 1. In VVSNet-R, residual connections with a kernel size of 3^3 are also added between transposed convolutional layers. The stride of these transposed convolution layers changes to 1 and there is an unpooling layer between two residual modules. The decoder generates a 32^3 volumetric grid finally.

C. Scaling Volume-view Supervision

Considering predicting more accurate reconstruction models, we proposed a scaling volume-view supervision module to build view guidance for the training. As shown in Fig.2, the module consists of an up-sampling layer, perspective transformation, projection operation, and a down-sampling layer. The size of the output volume after the up-sampling layer becomes 64. The purpose of this processing is that up-sampling can zoom the original reconstruction error and increase the view loss, and then improve the reconstruction accuracy. Given input 3D voxel and parameterized camera viewpoint, we can obtain 2D silhouettes via perspective transformation motivated by [7]. The formula can be written as follows:

$$U_i = \sum_n^H \sum_m^W \sum_l^D V_{nml} \max(1 - |x_i^s - m|, 0) \\ \max(1 - |y_i^s - n|, 0) \max(1 - |z_i^s - l|, 0) \\ \forall i \in [1...H^{'} W^{'} D^{'}]$$

where V_i represents the i-th voxel value. n, m, l is the n-th, m-th, and l-th pixel in height, width and depth, (x_i^s, y_i^s, z_i^s) is the coordinate of input volume V, (H, W, D) and $(H^{'}, W^{'}, D^{'})$ are the height, width, and depth of input volume V and output volume U. Then max operation is utilized to projection operation, because 3D voxel U is a binary unit, where 0 denotes an empty cell and 1 is a solid cell. The projection function as follows:

$$S_{n'm'} = \max_{l'} U_{n'm'l'}$$

then we obtain 64^2 2D silhouettes, the output is passed to a downsampling layer and compressed into 32^2 silhouettes, which are utilized to build the scaling volume-view supervision.

III. Optimization

The objective of the optimization is to enforce the 3D volumetric prediction and corresponding 2D silhouettes to resemble the ground truth volumetric occupancy grid and ground truth 2D masks, respectively. Therefore, the loss function of this framework includes two main parts: a volume reconstruction loss and a silhouette-based loss. The loss function is defined by mean squared error, so the loss in 3D space can be written as:

$$L_{vol} = \|V - V_{gt}\|^2$$

In addition, we consider that 2D silhouettes projected from the voxel occupancy grid should match the corresponding 2D ground truth masks well, then the voxel occupancy grid can be considered as a good match with the ground truth 3D volume. Therefore, the silhouette-based loss can be formulated as follows:

$$L_{sil} = \frac{1}{n} \sum_{i=1}^{n} \left\|S^i - S_{gt}^i\right\|^2$$

where i refers the index of 2D silhouette. Then the combination of both loss can be defined as:

$$L_{comb} = \lambda_{vol} L_{vol} + \lambda_{sil} L_{sil}$$

where λ_{vol} and λ_{sil} are weights that make different trade-offs between the volume reconstruction loss and silhouette-based loss. We evaluate this trade-off in detail in the experimental tests.

IV. Experiment

The dataset for training and testing the proposed framework and implementation details are introduced in this section. In addition, we conduct several experiments to analyze the impact of two version framework VVSNet and VVSNet-R. The influence of weighting factors λ_{vol} and λ_{sil} in the scaling volume-view supervision is evaluated. Moreover, the performance comparison of different methods is then verified. Finally, the generation ability to reconstruct unseen categories is evaluated in this section.

A. Dataset and Implementation Details

We evaluate the proposed VVSNet and VVSNet-R on both synthetic images from [7], which is based on the ShapeNet dataset [27]. More specifically, we use 13 major classes and 43,783 models.

To assess the quality of the 3D volume output, we use the intersection over union(IoU) between predicted 3D volume and corresponding ground truth as the similarity measure. The IoU is formally defined as follows:

$$IoU = \frac{\sum_{i=1}^{N} [I(y_i > t) * I(y_{gt})]}{\sum_{i=1}^{N} [I(I(y_i > t) + I(y_{gt}))]}$$

where y_i is the predicted value for the i-th voxel and y_{gt} is its corresponding ground truth value. $I(\cdot)$ is an indicator function, t refers a voxelization threshold, N denotes the total number of voxels. Higher IoU values indicate better reconstruction performance.

We implement the proposed network in Pytorch and train VVSNet and VVSNet-R using a batch size of 16 to fit in an NVIDIA Titan X GPU. We set Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$ and the learning rate is 0.001.

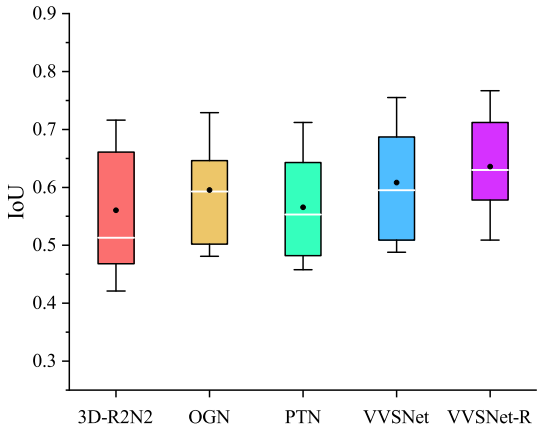

Fig. 3. Reconstruction performance is reported in mean (black dot) and median (white line) IoU value. The box plot shows 25% and 75% and caps show 15% and 85%.

B. Comparison of Different Methods

To validate the performance of the proposed framework VVSNet and VVSNet-R, we compared our methods with several state-of-the-art approaches including 3D-R2N2 [10], Octree Generating Network (OGN) [18] and Perspective Transformer Nets (PTN) [7] on the ShapeNet dataset. 3D-R2N2 utilizes 3D LSTM to fuse informative features and predict full 3D geometry from single or multiple views. OGN introduces a deep convolutional decoder architecture that can generate 3D shapes from a single view using an octree representation. PTN generates volumetric 3D outputs from a single view using perspective transformations. To make a fair comparison, we divided the dataset into a data set of 80% for the training set and 20% for the test set, which is the same as 3D-R2N2 and OGN. Moreover, we retrained the released PTN model using these categories.

As shown in Table 1 and Fig. 3, it can be seen that VVSNet and VVSNet-R significantly outperform other methods in all classes. In addition, VVSNet-R increases IoU over VVSNet by 14%. The visualization results are shown in Fig. 4, it can be found that VVSNet and VVSNet-R provide more complete volumetric 3D shapes, while other methods miss some part of detail regions, such as the lamp bracket and the table legs. It demonstrates that the scaling volume-view supervision makes a lot of contributions to the reconstruction results.

TABLE I
SINGLE-VIEW RECONSTRUCTION COMPARISON USING IOU.

Category	3D-R2N2	OGN	PTN	VVSNet	VVSNet-R
plane	0.513	0.587	0.553	0.599	**0.611**
bench	0.421	0.481	0.482	0.502	**0.509**
cabinet	0.716	0.729	0.711	0.755	**0.767**
chair	0.466	0.483	0.458	0.488	**0.509**
car	0.798	0.816	0.712	0.798	**0.816**
monitor	0.468	0.502	0.535	0.569	**0.578**
lamp	0.381	0.398	0.354	0.400	**0.411**
speaker	0.662	0.637	0.586	0.685	**0.712**
firearm	0.544	0.593	0.582	0.595	**0.612**
couch	0.628	0.646	0.643	0.687	**0.698**
table	0.513	0.536	0.471	0.623	**0.630**
cellphone	0.661	0.702	0.728	0.716	**0.762**
watercraft	0.513	0.632	0.536	0.583	**0.652**

TABLE II
IOU WHETHER USING UP-SAMPLING LAYER.

Model	IoU
VSSNet (no up-sampling layer)	0.591
VSSNet (with up-sampling layer)	0.599
VSSNet-R (no up-sampling layer)	0.602
VSSNet-R (with up-sampling layer)	0.611

C. Scaling Volume-view Supervision Evaluation

To demonstrate that designing up-sampling layers in the scaling volume-view supervision can increase errors and improve reconstruction accuracy, we compare the reconstruction performance whether the up-sampling layer is adopted. It worth noting that in order to ensure the consistency of resolution, the down-sampling layer is also not adopted when the up-sampling layer is dropped. For comparison, we trained these frameworks on the plane category. Table 2 and Fig. 5 show the quantitative results of IoU. We can see that with the help of the up-sampling layer, the IoU of VVSNet and VVSNet-R increase 1.3% and 1.4% respectively, which demonstrates that

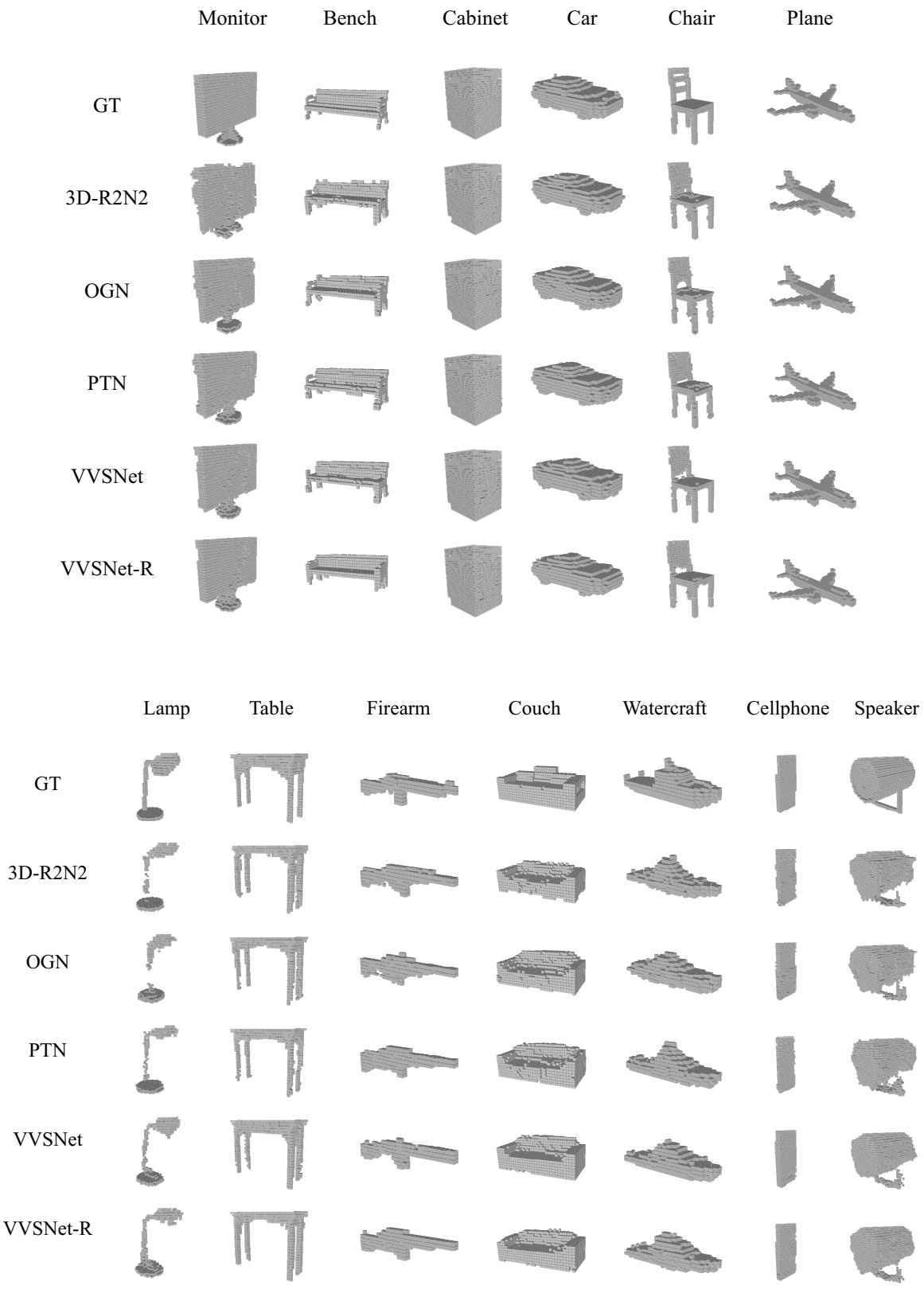

Fig. 4. Qualitative reconstruction results of the main 13 categories on the ShapeNet dataset compared with 3D-R2N2 [10], OGN [18] and PTN [7]. GT presents the ground truth of the objects. It shows VVSNet and VVSNet-R are better to make complete reconstruction with more details.

the up-sampling layer helps models to zoom errors and boost the reconstruction performance.

In addition, the performance of VVSNet-R is better than VVSNet overall. It can be seen that VVSNet-R using the deep residual module makes improvements in reconstruction performance.

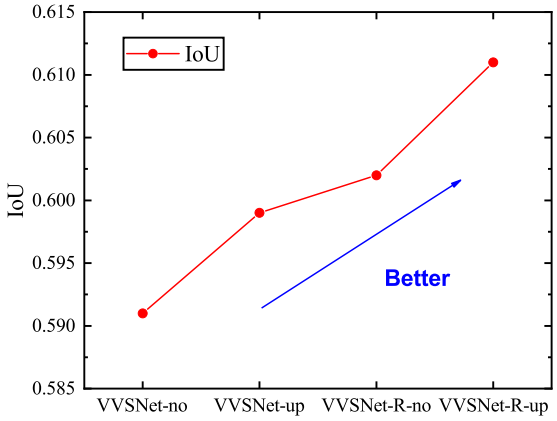

Fig. 5. Performance of VVSNet and VVSNet-R models with different supervision modules. VVSNet-no means VVSNet without the up-sampling layer and VVSNet-up presents VVSNet with the up-sampling layer.

D. Trade-off between View and Volume Supervision

To evaluate the contribution of the view and volume supervision in the optimization process, we trained the two proposed frameworks with volume supervision only, view supervision only, and volume-view supervision. For volume-view supervision, we set λ_{vol} and λ_{sil} are 0.5. We tested the performance on the couch category and Table 3 shows the prediction IoU. The use of both the volume and view supervision has achieved higher performance for both VVSNet and VVSNet-R overall. Without view loss, the performance of VVSNet and VVSNet-R decrease by 19% on average in IoU. In addition, Without volume loss, the performance of VVSNet and VVSNet-R decrease by 13% on average in IoU. We can conclude that the effect of view supervision is greater than that of volume supervision.

TABLE III
IoU USING DIFFERENT SUPERVISIONS.

	Encoder	Decoder	IoU
volume supervision	simple	simple	0.674
view supervision	simple	simple	0.676
volume-view supervision	simple	simple	0.687
volume supervision	residual	residual	0.686
view supervision	residual	residual	0.690
volume-view supervision	residual	residual	0.698

E. Evaluation of Generation

To evaluate the generation ability of the proposed methods, we leveraged our proposed methods to test several unseen categories: bed, cabinet, motorbike, train, and bookshelf that do not belong to 13 major classes. The visualization reconstruction results are shown in Fig. 6. We can see that the bed legs can not be reconstructed completely using the model VVSNet trained with volume supervision or view supervision only. However, the reconstruction performance gets better when the volume-view supervision is adopted. With the help of the deep residual module, VVSNet-R gets better performance compared with VVSNet. The bookshelf class is predicted most poorly because it is hard to find classes that are similar in shape to the training set. For the train, motorbike and car categories, the main parts are well reconstructed, but there are some missing details.

The quantitative analysis is shown in Table 4, it can be seen that the overall performance of the network is not good when only single supervision is used. The scaling volume-view supervision shows a powerful ability to handle the 3D reconstruction. In addition, the quantitative results demonstrate that the deep residual model improves performance.

V. Conclusion

In this paper, we propose the scaling volume-view supervision for an end-to-end 3D object reconstruction method. The scaling view supervision uses the up-sampling layer for zooming errors and increasing the penalties. The VVSNet provides a basic and flexible framework that uses 2D and 3D CNNs for efficient reconstruction learning. The VVSNet-R leverages the deep residual module to refine the general framework and boosts the prediction performance. We test our methods on the ShapeNet dataset and can see that the proposed methods get higher performance compared with state-of-the-art approaches. In addition, the scaling volume-view supervision makes great contributions to regularization and improves the reconstruction accuracy.

For future work, more experiments for real-world images will be conducted by using our methods. Moreover, we will introduce more flexible frameworks for both single and multiple-view reconstruction.

Acknowledgment

This work was supported by the National Key Research and Development Plan (2017YFC0806501) and National Natural Science Foundation (U1713224, 61973300).

TABLE IV
PERFORMANCE OF DIFFERENT VARIANT VVSNET ON THE UNSEEN CATEGORIES.

	bed	motorbike	bookshelf	train	bus
VVSNet-volume	0.113	0.166	0.077	0.179	0.112
VVSNet-view	0.147	0.197	0.163	0.201	0.156
VVSNet-view-volume	0.218	0.235	0.261	0.214	0.265
VVSNet-R-volume	0.187	0.226	0.279	0.281	0.306
VVSNet-R-view	0.256	0.385	0.301	0.309	0.318
VVSNet-R-view-volume	0.261	0.392	0.316	0.328	0.320

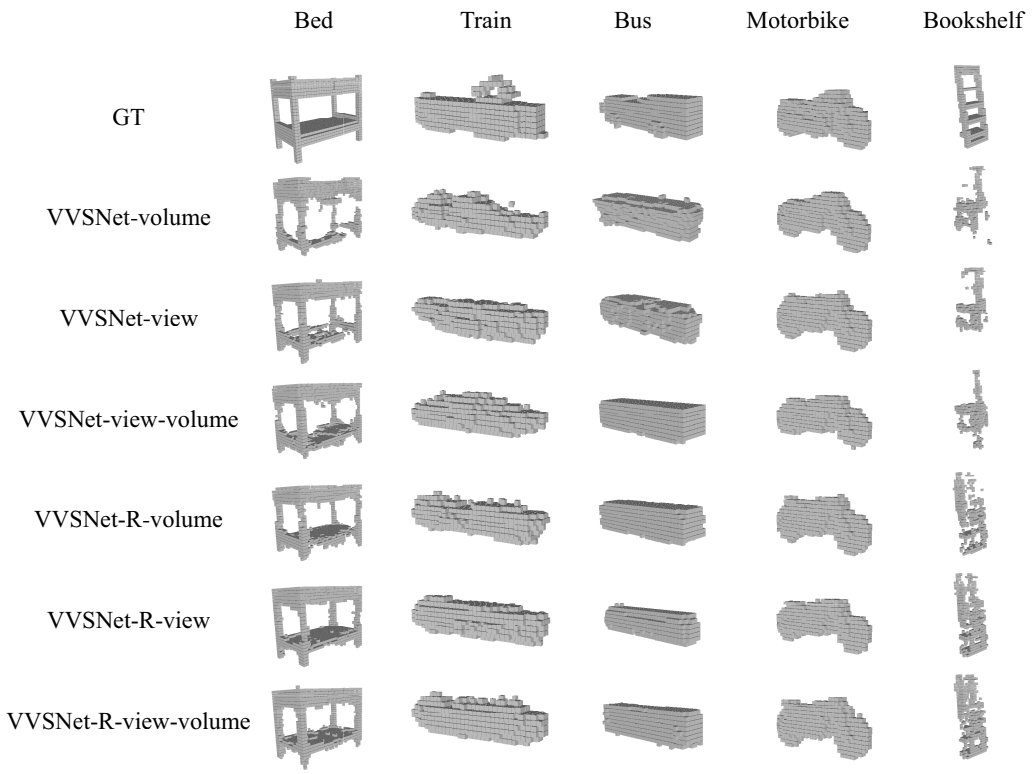

Fig. 6. 3D visualization reconstruction results on unseen categories. GT presents the ground truth of the objects. The scaling volume-view supervision and deep residual module boost the generation performance.

REFERENCES

[1] E. Dibra, H. Jain, C. Oztireli, *et al.*, "Human shape from silhouettes using generative hks descriptors and cross-modal neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4826–4836 (2017).

[2] K. Li, R. Garg, M. Cai, *et al.*, "Optimizable object reconstruction from a single view," *CoRR* **abs/1811.11921** (2018).

[3] J. T. Barron and J. Malik, "Shape, illumination, and reflectance from shading," *IEEE transactions on pattern analysis and machine intelligence* **37**(8), 1670–1687 (2014).

[4] S. R. Richter and S. Roth, "Discriminative shape from shading in uncalibrated illumination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1128–1136 (2015).

[5] A. P. Witkin, "Recovering surface shape and orientation from texture," *Artificial intelligence* **17**(1-3), 17–45 (1981).

[6] M. Tatarchenko, S. R. Richter, R. Ranftl, *et al.*, "What do single-view 3d reconstruction networks learn?," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3405–3414 (2019).

[7] X. Yan, J. Yang, E. Yumer, *et al.*, "Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision," in *Advances in Neural Information Processing Systems*, 1696–1704 (2016).

[8] A. M. Terwilliger, G. N. Perdue, D. Isele, *et al.*, "Vertex reconstruction of neutrino interactions using deep learning," in *2017 International Joint Conference on Neural Networks (IJCNN)*, 2275–2281 (2017).

[9] A. Kar, C. Häne, and J. Malik, "Learning a multi-view stereo machine," in *Advances in neural information processing systems*, 365–376 (2017).

[10] C. B. Choy, D. Xu, J. Gwak, *et al.*, "3d-r2n2: A unified approach for single and multi-view 3d object reconstruction," in *European conference on computer vision*, 628–644, Springer (2016).

[11] A. Johnston, R. Garg, G. Carneiro, *et al.*, "Scaling cnns for high resolution volumetric reconstruction from a single image," in *Proceedings of the IEEE International Conference on Computer Vision*, 939–948 (2017).

[12] X. Yang, Y. Wang, Y. Wang, *et al.*, "Active object reconstruction using a guided view planner," *arXiv preprint arXiv:1805.03081* (2018).

[13] X. Li, Y. Dong, P. Peers, *et al.*, "Synthesizing 3d shapes from sil-

houette image collections using multi-projection generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5535–5544 (2019).

[14] B. Yang, H. Wen, S. Wang, *et al.*, "3d object reconstruction from a single depth view with adversarial learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 679–688 (2017).

[15] J. Wu, C. Zhang, T. Xue, *et al.*, "Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling," in *Advances in neural information processing systems*, 82–90 (2016).

[16] L. Jiang, S. Shi, X. Qi, *et al.*, "Gal: Geometric adversarial loss for single-view 3d-object reconstruction," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 802–816 (2018).

[17] P.-S. Wang, Y. Liu, Y.-X. Guo, *et al.*, "O-cnn: Octree-based convolutional neural networks for 3d shape analysis," *ACM Transactions on Graphics (TOG)* **36**(4), 72 (2017).

[18] M. Tatarchenko, A. Dosovitskiy, and T. Brox, "Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs," in *Proceedings of the IEEE International Conference on Computer Vision*, 2088–2096 (2017).

[19] H. Kato, Y. Ushiku, and T. Harada, "Neural 3d mesh renderer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3907–3916 (2018).

[20] N. Wang, Y. Zhang, Z. Li, *et al.*, "Pixel2mesh: Generating 3d mesh models from single rgb images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 52–67 (2018).

[21] J. K. Pontes, C. Kong, S. Sridharan, *et al.*, "Image2mesh: A learning framework for single image 3d reconstruction," in *Asian Conference on Computer Vision*, 365–381, Springer (2018).

[22] P. Mandikal and V. B. Radhakrishnan, "Dense 3d point cloud reconstruction using a deep pyramid network," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1052–1060, IEEE (2019).

[23] P. Mandikal, N. Murthy, M. Agarwal, *et al.*, "3d-lmnet: Latent embedding matching for accurate and diverse 3d point cloud reconstruction from a single image," *arXiv preprint arXiv:1807.07796* (2018).

[24] H. Fan, H. Su, and L. J. Guibas, "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*, 605–613 (2017).

[25] M. Ye, S. Zhou, and H. Fu, "Deepshapesketch : Generating hand drawing sketches from 3d objects," in *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–8 (2019).

[26] J. Wu, T. Xue, J. J. Lim, *et al.*, "Single image 3d interpreter network,"

[27] Z. Wu, S. Song, A. Khosla, *et al.*, "3d shapenets: A deep representation for volumetric shapes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1912–1920 (2015).