IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON MOBILE MULTIMEDIA: METHODOLOGY AND APPLICATIONS

Received December 25, 2019, accepted January 17, 2020, date of publication January 22, 2020, date of current version January 29, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2968489

Deep Auxiliary Learning for Point Cloud Generation

FEI HU, LONG YE, WEI ZHONG, LI FANG, AND QIN ZHANG

Key Laboratory of Media Audio and Video, Communication University of China, Ministry of Education, Beijing 100024, China

Corresponding author: Long Ye (yelong@cuc.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61971383, Grant 61631016, and Grant 61801440.

ABSTRACT Generation point cloud from single image is a classical problem in computer vision. The learning methods for this task often adopt local distance metrics as loss function, which means the generated points are not easy to meet the overall shape distribution of the target object. To solve this problem, we introduce a voxel reconstruction network with distribution fitting as auxiliary task and propose a novel framework named Voxel-Assisted Points Generation Network(VAPGN). The auxiliary learning with voxel generation makes it easier to capture the shape distribution of objects in the image during the encoder phase, thereby effectively improving the result of point cloud reconstruction. To meet the needs of mobile and embedded applications, a mobile version of the model is also proposed. In the experiments, we verify the feasibility of our network on the ShapeNet dataset. The proposed framework has achieved outstanding performance on the point cloud generation task, comparing with various state-of-the-art methods.

INDEX TERMS Point cloud generation, auxiliary learning.

I. INTRODUCTION

The task of 3D reconstruction from one or several 2D images is a classic issue in computer vision [1]–[5]. It is a generally scientific problem in a wide variety of fields, such as virtual reality [6]–[8], autopilot [9], [10], etc. One of the key difficulties is how to effectively model the 3D information in the image.

By observing a 2D image, humans can easily perceive the 3D information in it. At the moment of seeing a picture, humans can recognize the target contained therein, the 2D perspective projection shape of the target, and 3D shapes formed by the target under certain lighting conditions, thereby restore the 3D information. That is to say, the picture itself is not complete for perceiving the 3D information of the picture, and human intelligence needs to use the summarized rules and common sense to make inferences. This means that we can complete 3D reconstruction in a data-driven way. The emergence of large-scale 3D shape datasets [11], [12] and the development of data-driven methods [13], [14] have given rise to new interest in reconstructing 3D shape with learning method.

The associate editor coordinating the review of this manuscript and approving it for publication was Dapeng Wu.

As high resolution voxel grids require huge memory, many works focus on the point cloud generation. These works often adopt distance metrics(earth mover's distance, chamfer distance etc.) as loss function, which means the generated points are not easy to meet the overall shape distribution of the target object. However, the voxel generation model usually targets distribution fitting so that the generated volume matches the shape of the real object well. Therefore, the voxel generation can be adopted as a suitable auxiliary task to supplement the defects of the point cloud generation model.

Motivated by this idea, we propose a novel architecture to finish the image-based point cloud reconstruction, referred to as Voxel-Assisted Points Generation Network(VAPGN). The proposed model consists of two neural networks, one of which is the main network to finish the point generation task, and the other is the auxiliary learning network whose outputs are voxel grids. These two tasks share some convolutional layers in learning process to learn the general feature from input images for 3D reconstruction, and the exactly reconstruction process from the feature vector with separate decoder branches. The whole network is trained by a combined loss function. Another interesting aspect of our architecture is that we design the encoder with a pretrained image classification model.

To meet the needs of mobile and embedded applications, we make some corresponding alterations to the original model, and proposes a mobile version of the proposed model.

The main contributions of our work are summarized below:

- We propose a auxiliary approach to model the intrinsic general 3D characteristics in point cloud reconstruction. The auxiliary learning of voxel generation makes it easier to capture the shape distribution of objects in the image during the encoder phase of the point cloud reconstruction task, thereby effectively improving the result of point cloud reconstruction.
- We propose a mobile version of the proposed model for mobile applications
- We propose a reasonable combined loss function for our framework.
- Our extensive experiment demonstrates the outstanding performance of our reconstruction algorithm compared to the state-of-the-art techniques on the public dataset ShapeNet.

The rest of the paper is organised as follows. Section 2 presents related literature. Section 3 illustrates the proposed network architecture, training losses and training strategy. Section 4 presents experiments, ablation study and analysis. Finally, we provide a brief summarisation of our work in Section 5.

II. RELATED WORKS

A. DEEP LEARNING ON 3D RECONSTRUCTION

Existing work 3D reconstruction bases on learning can be roughly classified as voxels based, point cloud based or surface based according to the output representations they produce.

1) VOXEL

A voxel is an abbreviation for a volume element. Due to the simplicity, voxels are the most commonly used representation for 3D tasks. Wu *et al.* [11] combined 2.5D depth map and 3D shape class to complete the reconstruction task. Girdhar *et al.* [15] fed the image and 3D voxel grid into their TL-embedding network to train a predictable vector for 3D object reconstruction. Yan *et al.* [16] proposed an encoder-decoder network with a projection loss defined by the perspective transformation which enables the unsupervised learning using 2D observation and fixed-viewpoint without explicit 3D supervision. Wu *et al.* [17] proposed 3D Generative Adversarial Network which generates 3D objects from a probabilistic space. This research first introduced GAN to 3D field.

The issue of reconstructing 3D geometry from multiple views has been considered in [18]–[20]. Choy *et al.* [18] proposed an overall framework called 3D Recurrent Reconstruction Neural Network for single-view and multi-view reconstruction based on LSTM which means it had high computation complexity. Ji *et al.* [19] and Kar *et al.* [20] encode camera parameters with the input image as a 3D voxel representation and apply a 3D convolution to reconstruct the 3D scene from multiple views.

However, due to memory limitation, these methods are limited to relatively small 32^3 resolutions. Although recent work has applied 3D convolutional neural networks to resolutions as high as 128^3, this only applies to small batch sizes, which results in slow training.

Due to the high memory requirements expressed by voxels, recent work has proposed to reconstruct 3D objects in a multi-resolution manner. However, the resulting method is often difficult to implement and requires multiple passes on the input to generate the final 3D model. In addition, they are still limited to 256^3 voxel grids.

2) POINT CLOUD

The point cloud has also been widely used in computer graphics. Qi *et al.* [21], [22] pioneered the point cloud as a manifestation of discriminative deep learning tasks. They achieve alignment invariance by applying a fully connected neural network to each point independently and then performing global pooling operations. Fan *et al.* [23] proposed a point set generation network for 3D object reconstruction from one single image. Although named point cloud, it actually learns the coordinates of the voxel grid. And the fixed number of points would not do any help to represent complex geometric structure of objects.

By applying a convolution on the graph spanned by the vertices and edges of the grid, the grid is first considered for discriminative 3D classification tasks [24], [25]. Recently, the grid has also been considered to be the output representation of 3D reconstruction [26]–[28]. Liao *et al.* [27] proposed an end-to-end readable version of the marching cube algorithm. However, their approach is still limited by the memory requirements of the underlying 3D mesh and is therefore limited to 32^3 voxel resolution.

3) SURFACE

There are also many implicit surfaces representation methods based on deep learning [29]–[32].Park *et al.* [32] represented a shape's surface by a continuous volumetric field. Michalkiewicz *et al.* [31] proposed an end-to-end trainable model that directly predicts implicit surface representations of arbitrary topology by optimising a novel geometric loss function.

B. MULTI-TASK LEARNING

The proposed approach belongs to multi-task learning (MTL), which refers to a learning paradigm in machine learning which aims to leverage useful information contained in multiple related tasks to help improve the generalization performance of some tasks.

Argyriou *et al.* [33] presented a method for learning sparse representations shared across multiple tasks. Popa *et al.* [34] proposed a deep multitask architecture for fully automatic 2d and 3d human sensing (DMHS), including recognition and reconstruction, in monocular images. Ranjan *et al.* [35]

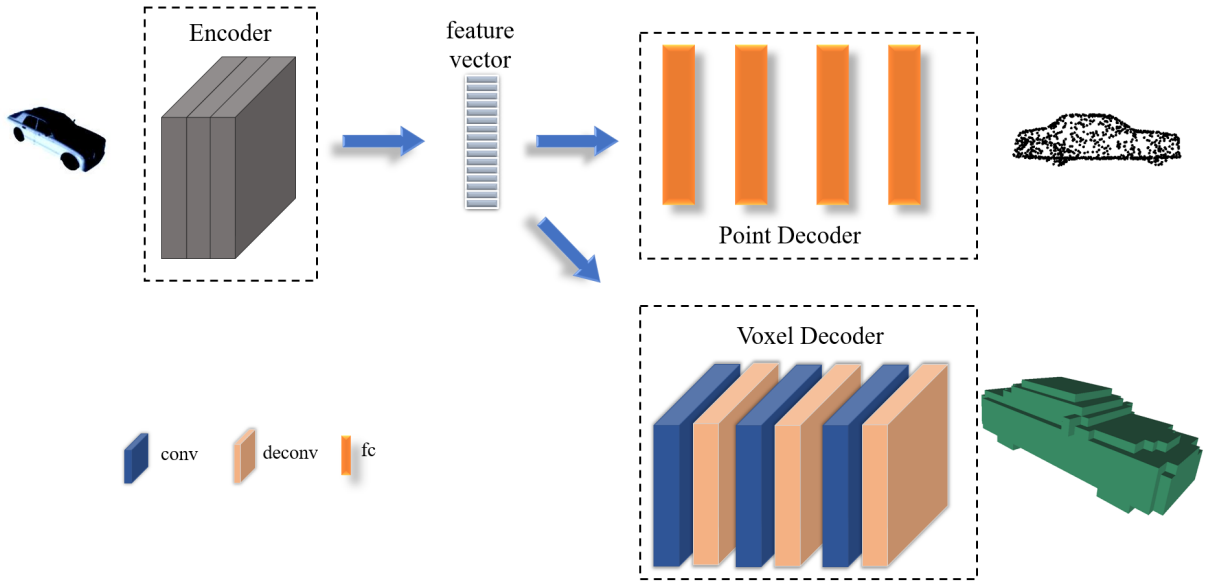

FIGURE 1. The framework of DDRN.

proposed two novel CNN architectures that perform face detection, landmarks localization, pose estimation and gender recognition by fusing the intermediate layers of the network. Ma *et al.* [36] proposed a multi-task end-to-end optimized deep neural network (MEON) for blind image quality assessment. Liu *et al.* [37] proposed a hierarchical clustering multi-task learning (HC-MTL) method for joint human action grouping and recognition. Ke *et al.* [38] proposed to use deep convolutional neural networks to learn long-term temporal information of the skeleton sequence from the frames of the generated clips, and then used a Multi-Task Learning Network (MTLN) to jointly process all frames of the clips in parallel to incorporate spatial structural information for action recognition. Dou *et al.* [39] divided the 3D face reconstruction into neutral 3D facial shape reconstruction and expressive 3D facial shape reconstruction and trained the combine model with a multi-task loss function.

III. VAPGN

This section describes our framework architecture and related loss function.

A. NETWORK ARCHITECTURE

Figure 1 shows the overall structure of the proposed method, starting from a single image and ending with different predictions. We use two networks performing the point cloud generation and voxel generation tasks by sharing features, which forms a tree-structured network architecture. Both tasks adopt the encoder-decoder structure. The two networks share a set of encoder layers for extracting the features of a given image on multiple levels, and meanwhile have different decoder layers that are tailored for different applications, including point cloud generation and voxel generation.

According to the theory in [40], the current state-of-the-art in single-view object reconstruction does not actually perform reconstruction but image classification. The convolutional layers of the encoder are identical to the corresponding parts of state-of-the-art classification model ResNet34 except that the feature map sizes are adjusted by our input size. Due to the limitation of computing power, we have not adopted a more complex network structure such as Res101, Inception etc. We also compress the network layers as much as possible in decoder layers to reduce the parameters. The voxel decoder consists of only three convolutional layers and three deconvolutional layers, and the point decoder consists of only four fully connected layers.

In the point generation task, the feature vector extracted from the shared layers are fed into the point decoder layers to produce point cloud of the object in the input image. To make the output map have the same size as the ground truth, we design a series of four fully connected layers. The task finally outputs a $1024 * 3$ tensor, representing a series of 1024 point cloud coordinates. We can visualize the point cloud according to these coordinates.

In the voxel generation task, we aim to generate a $32 * 32 * 32$ voxel grid. Specifically, we use three convolutional layers and three deconvolutional layers. The task finally outputs a $32 * 32 * 32$ tensor. We then convert the output vector with voxel-wise sigmoid function and transform it into voxel occupancy through thresholding.

The whole model receives a $224 * 224$ RGB input, converts it into a feature containing high level information, then completes the respective task with individual branch. Note that we refer to voxel decoder and point decoder as domain-specific layers and all the feature preceding layers as shared layers, the feature encoder extracts generic features for 3D reconstruction with joint learning.

TABLE 1. The Chamfer-L1 score on the ShapeNet data set.The lower the score, the better the performance.

category	3D-R2N2	PSGN	Pix2Mesh	AtlasNet	Occupancy	Ours ($w = 20$)	Ours ($w = 10$)
airplane	0.227	0.137	0.187	0.104	0.147	**0.0213**	0.0244
bench	0.194	0.181	0.201	0.138	0.155	**0.0272**	0.0301
cabinet	0.217	0.215	0.196	0.175	0.167	**0.0309**	0.0342
car	0.213	0.169	0.18	0.141	0.159	**0.0245**	0.0265
chair	0.27	0.247	0.265	0.209	0.228	**0.0357**	0.0396
display	0.314	0.284	0.239	0.198	0.278	**0.0377**	0.0415
lamp	0.778	0.314	0.308	0.305	0.479	**0.045**	0.0479
loudspeaker	0.318	0.316	0.285	0.245	0.3	**0.0425**	0.0468
rifle	0.183	0.134	0.164	0.115	0.141	**0.0197**	0.0214
sofa	0.229	0.224	0.212	0.177	0.194	**0.0316**	0.037
table	0.239	0.222	0.218	0.19	0.189	**0.0318**	0.0351
telephone	0.195	0.161	0.149	0.128	0.14	**0.0235**	0.0276
vessel	0.238	0.188	0.212	0.151	0.218	**0.0276**	0.0299
mean	0.278	0.215	0.216	0.175	0.215	**0.0307**	0.034

B. METRICS

The Chamfer-L1 distance is defined as the mean of accuracy and a completeness metric [30]. The accuracy metric is defined as the mean L1 distance of points on the output mesh to their nearest neighbors on the ground truth mesh. The completeness metric is defined similarly, but in opposite direction. The corresponding distances are estimated with a KD-tree. The Chamfer-L2 distance is defined similarly to the Chamfer-L1 distance.

Intersection over Union (IoU), also known as the Jaccard index, is the most popular evaluation metric for tasks such as segmentation, object detection and tracking. The IoU in our task is defined as the division of two volume grid sets' intersection by the union of the two.

C. LOSS FUNCTION

Loss function is a critical factor for the convergence of neural network. A combined loss function has been introduced in this task.

Let $I = \{I_i\}_{i=1}^{N_1}$ denote a collection of training images, $V = \{V_{i1,i2,i3,i4} \in \{0,1\}\}_{N_1\times 32\times 32\times 32}$ denote their corresponding element-wise ground truth voxel occupancies, and $P = \{P_j | P_j \in R_+^3\}_{N_1 * N_2}$ denote their corresponding point-wise ground truth coordinates. Furthermore, we denote all the parameters in the shared layers as θ_s; the parameters in the point cloud generation task as θ_p, and the parameters in the voxel generation task as θ_v.

$$\begin{aligned} L_v(I;\theta_s,\theta_v) = -\frac{1}{N_1\times 32\times 32\times 32}\sum_{i1=1}^{N_1}\sum_{i=0}^{1}\sum_{i2=1}^{32}\sum_{i3=1}^{32}\sum_{i4=1}^{32} \\ \times[1\{V_{i1,i2,i3,i4}=i\}\log(h_{i1,i,i2,i3,i4}(I_{i1};\theta_s,\theta_v)) \\ +(1-1\{V_{i1,i2,i3,i4}=i\}) \\ *(1-\log(h_{i1,i,i2,i3,i4}(I_{i1};\theta_s,\theta_v)))], \end{aligned} \tag{1}$$

$$\begin{aligned} L_p(I;\theta_s,\theta_p) = \sum_{k=1}^{N_1}(\sum_{i=1}^{N_3}\min_j ||p_{k,j}-r_{k,i}||_2^2 \\ +\sum_{j=1}^{N_2}\min_i ||r_{k,i}-p_{k,j}||_2^2), \end{aligned} \tag{2}$$

where $1\{\cdot\}$ is the indicator function, h is the voxel generation function, and $h_{i1,i,i2,i3,i4}$ is the $(i2, i3, i4)$-th element of the $i1$-th probabilistic voxel occupancy map; $r_{k,i}$ is the k-th point cloud gotten from the i-th imager with the point generation function. Clearly, the first one is associated with the binary cross entropy loss term for the voxel generation task, while the second function corresponds to the Chamfer-L2 loss term for the point generation task. Then our optimization problem is

$$\begin{aligned} &\text{minimize } L_p(I;\theta_s,\theta_p) \\ &\text{subject to } L_v(I;\theta_s,\theta_v) < \epsilon, \end{aligned} \tag{3}$$

where ϵ is a small positive number.Suppose the optimal value of the problem is v^* According to our daily prior knowledge, there is a great correlation between the two generation parts, so we assume that there is a function g that satisfies $\theta_v = g(\theta_p)$.The problem can be transformed into:

$$\begin{aligned} &\text{minimize } L_p(I;\theta_s,\theta_p) \\ &\text{subject to } L_v(I;\theta_s,g(\theta_p)) < \epsilon, \end{aligned} \tag{4}$$

The Lagrange dual function of this problem is:

$$L_d(\theta_s,\theta_p,\lambda) = L_p(I;\theta_s,\theta_p) + \lambda(L_v(I;\theta_s,g(\theta_p)) - \epsilon),$$

So,

$$g(\lambda) = \inf_{\theta_s,\theta_p} L_d(\theta_s,\theta_p,\lambda) \le v^*, \tag{5}$$

TABLE 2. The detail Chamfer-L1 score on the ShapeNet data set. Here we set $w = 20$ in our VAPGN according to table 1.

category	Accuracy			Completeness			Chamfer-L1		
	PSGN	ONet	Ours	PSGN	ONet	Ours	PSGN	ONet	Ours
airplane	0.113	0.133	0.0277	0.191	0.161	0.0149	0.152	0.147	0.0213
bench	0.161	0.154	0.0359	0.262	0.156	0.0185	0.212	0.155	0.0272
cabinet	0.154	0.15	0.0436	0.307	0.184	0.0181	0.231	0.167	0.0309
car	0.126	0.116	0.0331	0.235	0.203	0.0159	0.181	0.159	0.0245
chair	0.233	0.223	0.0467	0.333	0.233	0.0248	0.283	0.228	0.0357
display	0.258	0.281	0.0468	0.351	0.275	0.0287	0.304	0.278	0.0377
lamp	0.301	0.402	0.0522	0.372	0.557	0.0378	0.337	0.479	0.045
loudspeaker	0.253	0.285	0.0558	0.403	0.315	0.0292	0.328	0.3	0.0425
rifle	0.113	0.148	0.0247	0.159	0.134	0.0148	0.136	0.141	0.0197
sofa	0.206	0.194	0.0417	0.307	0.195	0.0214	0.257	0.194	0.0316
table	0.165	0.189	0.0446	0.307	0.189	0.0191	0.236	0.189	0.0318
telephone	0.14	0.149	0.0312	0.219	0.13	0.0158	0.179	0.14	0.0235
vessel	0.186	0.197	0.0339	0.249	0.238	0.0212	0.217	0.218	0.0276
mean	0.185	0.202	0.0398	0.284	0.228	0.0215	0.235	0.215	0.0307

This means that $g(\lambda)$ can give a lower bound on the objective problem with any given λ. If λ, ϵ are given, then

$$L_d(\theta_s, \theta_p, \lambda) = L_p(I; \theta_s, \theta_p) + \lambda(L_v(I; \theta_s, g(\theta_p))) + c,$$

where c is a constant number. So the proposed DDRN is trained by minimizing the following loss function L:

$$L = L_p(I; \theta_s, \theta_v) + w * L_v(I; \theta_s, \theta_p), \tag{6}$$

where w is the hyperparameter.

IV. EXPERIMENT

In this section, we firstly introduce the dataset and training details, and verify the feasibility of our algorithm in quantitatively (summarized in the Table 3 and Table 1) and qualitatively (shown in Figure 2).

Dataset: The ShapeNet dataset is a richly-annotated, large-scale dataset of 3D shapes which is collected by Princeton, Stanford and TTIC. We used a subset of the ShapeNet dataset which contains about 50,000 3D models over 13 common categories. We split the dataset into training, valid sets and test sets, with 50 percent for training, 30 percent for validation and the remaining 20 percent for testing. Note that all viewpoints are sampled randomly.

Implement Detail: We use the ADAM [41] solver for stochastic optimization in all the experiments. During the training process, we set the learning rate 10e-4 for the neural networks.

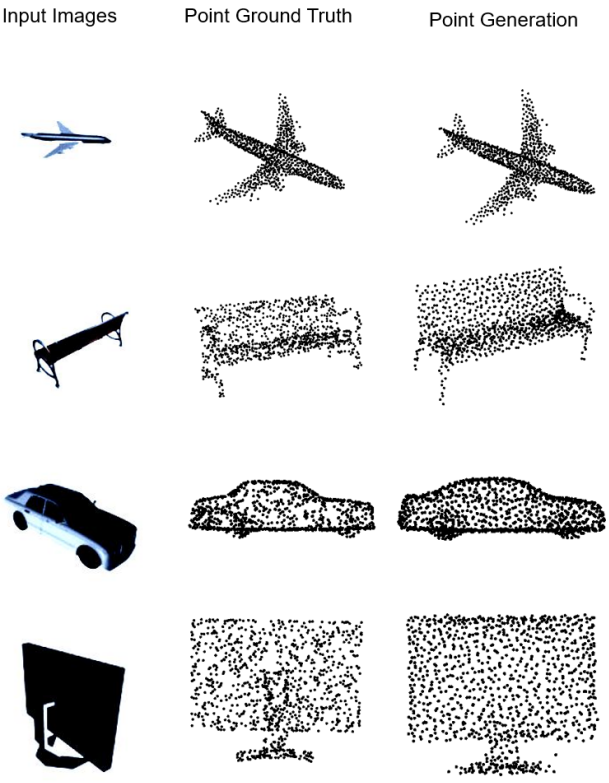

FIGURE 2. Some qualitative result instances. The input image is shown in the first column, the other columns show the result of our method compared with the ground truth.

A. STATE-OF-THE-ART PERFORMANCE COMPARISON

We compare the proposed approach with several state-of-the-art methods including 3D-R2N2 [18],PSGN [23], Pix2Mesh [28], AtlasNet [26], Occupancy Network [30].

Table 1 shows a numerical comparison of our approach and the state-of-the-art for single image point cloud reconstruction on the ShapeNet dataset. Our method achieves the highest Chamfer-L1 score to the ground truth.According to Table 1, we have achieved the best Chamfer-L1 score in all categories and almost improved by an order of magnitude. (As [30] reproduces some experiments and gets better performance, parts of the results are from [30] instead of the

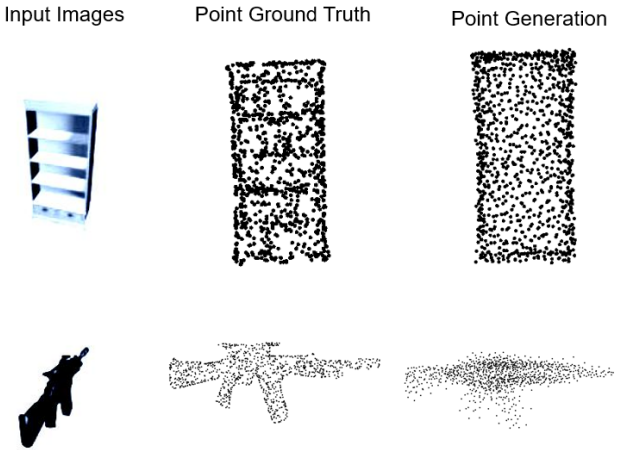

FIGURE 3. Some failure instances. The input image is shown in the first column, the second column is the ground truth, the other column shows the results of our method.

TABLE 3. The IoU score of voxel generation task on the ShapeNet data set.The higher the score, the better the performance.Since occupancy network does not provide the commonly used IoU scores, we download the author's pre-trained model from the official site provided by the author and obtain the corresponding IoU score.

	3D-R2N2	PSGN	Occupancy	Ours(w=20)
Plane	0.513	0.601	0.468	**0.621**
Bench	0.421	0.55	0.522	**0.564**
Cabinet	0.716	0.771	0.747	**0.779**
Car	0.798	0.831	0.784	**0.848**
Chair	0.466	0.544	0.553	**0.561**
monitor	0.468	0.552	**0.594**	0.58
Lamp	0.381	**0.462**	0.38	0.436
speaker	0.662	**0.737**	0.712	0.72
Firearm	0.544	**0.604**	0.516	0.596
Couch	0.628	0.708	**0.717**	0.713
Table	0.513	**0.606**	0.544	0.6
cellphone	0.661	0.749	0.742	**0.767**
watercraft	0.513	0.611	0.575	**0.623**
Mean	0.56	0.64	0.604	**0.647**

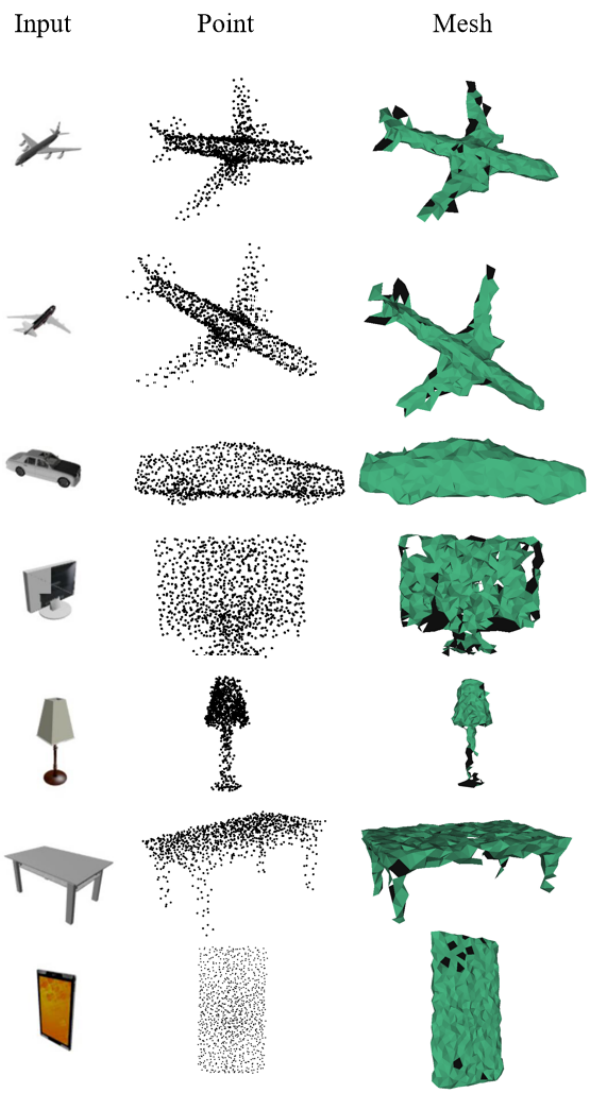

FIGURE 4. Some qualitative result instances of the mobile version. The input image is shown in the first column, the second columns show the result of our method, and we also reconstruction the surface with the algorithm from [42] in the third column.

original reference.) Some results from our model are visualized in Figure 2. Table 2 shows a more detailed Chamfer-L1 score, and Figure 3 shows a visual display of some failed instances.

Besides, the category-wise IoU score on voxel generation task is shown in Table 3, and we have achieved the best in 7 out of 13 categories and got the best mean score.

B. ABLATION STUDY

In order to verify our strategy, we evaluate the performance differences of the proposed approach with and without auxiliary learning. The proposed model without auxiliary task consists of the shared encoder and task-related decoder in Figure 1, resulting in a single-task version. The loss function for the single-task version is the Charmfer-L2 distance between the prediction and the ground truth as shown in Equation (2). Table 4 shows the quantitative comparison on the ShapeNet without or with auxiliary training, respectively. Clearly, auxiliary task helps to learn a better model for point cloud generation task, as performing voxel task improves the performance of point cloud generation in all the categories.

C. THE MOBILE VERSION

In real life, it is often necessary to consider implementing point cloud reconstruction tasks on computationally limited platforms such as robots, mobile phones etc. This section proposes a mobile version of the proposed model to meet the needs of mobile and embedded applications. This mobile version makes the following corresponding alterations to the original.

TABLE 4. The Chamfer-L1 score on the point cloud generation task of auxiliary learning(VAPGN) and single-task learning.

	VAPGN	Single-task
Plane	0.0213	0.121
Bench	0.0272	0.171
Cabinet	0.0309	0.245
Car	0.0245	0.170
Chair	0.0357	0.219
monitor	0.0377	0.210
Lamp	0.0450	0.231
speaker	0.0425	0.283
Firearm	0.0197	0.111
Couch	0.0316	0.198
Table	0.0318	0.298
cellphone	0.0235	0.161
watercraft	0.0276	0.128
Mean	0.0307	0.196

- Replace the original ResNet34 with MobileNets which is designed for mobile phone.
- Replace the original fully connected layer in the point cloud generation task with a convolutional layer of kernel size 1.

With these adjustments, the parameter amount can be successfully reduced from 29 million to 8 million. The Chamfer-L1 score on ShapeNet of our mobile vision is 0.068. Although the mobile version scores lower than the original version, it is still higher than the other networks mentioned above. Figure 4 shows some instances of the mobile model.

V. CONCLUSION

In this paper, we propose a simple yet effective auxiliary learning approach for point cloud reconstruction. The auxiliary structure helps the main network to learn better feature for point cloud generation. We also design a combined loss function for the model. A mobile version of the proposed model is proposed to meet the needs of mobile and embedded applications. Compared to the state-of-the-art, we have achieved outstanding performance in the large public 3D reconstruction dataset ShapeNet 3D+.

REFERENCES

[1] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski, "Building Rome in a day," *Commun. ACM*, vol. 54, no. 10, pp. 105–112, Oct. 2011.
[2] A. Criminisi, I. Reid, and A. Zisserman, "Single view metrology," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 123–148, 2000.
[3] G. E. Elsinga, F. Scarano, B. Wieneke, and B. W. Van Oudheusden, "Tomographic particle image velocimetry," *Exp. Fluids*, vol. 41, no. 6, pp. 933–947, 2006.
[4] M. Pollefeys, R. Koch, and L. Van Gool, "Self-calibration and metric reconstruction inspite of varying and unknown intrinsic camera parameters," *Int. J. Comput. Vis.*, vol. 32, no. 1, pp. 7–25, 1999.
[5] F. Remondino and S. El-Hakim, "Image-based 3D modelling: A review," *Photogramm. Rec.*, vol. 21, no. 115, pp. 269–291, Jul. 2010.
[6] F. Bruno, S. Bruno, G. De Sensi, M.-L. Luchi, S. Mancuso, and M. Muzzupappa, "From 3D reconstruction to virtual reality: A complete methodology for digital archaeological exhibition," *J. Cultural*, vol. 11, no. 1, pp. 42–49, Jan. 2010.
[7] S. El-Hakim, J.-A. Beraldin, M. Picard, and G. Godin, "Detailed 3D reconstruction of large-scale heritage sites with integrated techniques," *IEEE Comput. Grap. Appl.*, vol. 24, no. 3, pp. 21–29, May 2004.
[8] V. Sequeira, K. Ng, E. Wolfart, J. Gonçalves, and D. Hogg, "Automated reconstruction of 3D models from real environments," *ISPRS J. Photogram. Remote Sens.*, vol. 54, no. 1, pp. 1–22, Feb. 1999.
[9] A. Rozantsev, S. N. Sinha, D. Dey, and P. Fua, "Flight dynamics-based recovery of a UAV trajectory using ground cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6030–6039.
[10] X. Zheng, F. Wang, and Z. Li, "A multi-UAV cooperative route planning methodology for 3D fine-resolution building model reconstruction," *ISPRS J. Photogram. Remote Sens.*, vol. 146, pp. 483–494, Dec. 2018.
[11] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3D ShapeNets: A deep representation for volumetric shapes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015.
[12] Y. Xiang, W. Kim, W. Chen, J. Ji, C. Choy, H. Su, R. Mottaghi, L. Guibas, and S. Savarese, "ObjectNet3D: A large scale database for 3D object recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 160–176.
[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.
[14] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell. (AAAI)*, 2016, pp. 4278–4284.
[15] R. Girdhar, D. F. Fouhey, M. Rodriguez, and A. Gupta, "Learning a predictable and generative vector representation for objects," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 484–499.
[16] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee, "Perspective transformer nets: Learning single-view 3D object reconstruction without 3D supervision," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1696–1704.
[17] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum, "Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 82–90.
[18] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3D-R2N2: A unified approach for single and multi-view 3D object reconstruction," in *Proc. ECCV*, 2016, pp. 628–644.
[19] M. Ji, J. Gall, H. Zheng, Y. Liu, and L. Fang, "SurfaceNet: An end-to-end 3D neural network for multiview stereopsis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017.
[20] A. Kar, C. Hane, and J. Malik, "Learning a multi-view stereo machine," Aug. 2017, *arXiv:1708.05375*. [Online]. Available: https://arxiv.org/abs/1708.05375
[21] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017.
[22] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, 2017, pp. 1–10.
[23] H. Fan, H. Su, and L. Guibas, "A point set generation network for 3D object reconstruction from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 28, Jul. 2017.
[24] M. M. Bronstein, J. Bruna, Y. Lecun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: Going beyond Euclidean data," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 18–42, Jul. 2017.
[25] K. Guo, D. Zou, and X. Chen, "3D mesh labeling via deep convolutional neural networks," *TOGACM Trans. Graph.*, vol. 35, no. 1, pp. 1–12, Dec. 2015.
[26] T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry, "A papier-mache approach to learning 3D surface generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018.
[27] Y. Liao, S. Donne, and A. Geiger, "Deep marching cubes: Learning explicit surface representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018.
[28] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y. Jiang, "Pixel2Mesh: Generating 3D mesh models from single RGB images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 55–71.
[29] Z. Chen and H. Zhang, "Learning implicit fields for generative shape modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5939–5948.
[30] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3D reconstruction in function space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019.

[31] M. Michalkiewicz, J. K. Pontes, D. Jack, M. Baktashmotlagh, and A. Eriksson, "Deep level sets: Implicit surface representations for 3D shape inference," in *Proc. Comput. Vis. Pattern Recognit.*, 2019, pp. 1–10.
[32] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "DeepSDF: Learning continuous signed distance functions for shape representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019.
[33] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Mach. Learn.*, vol. 73, no. 3, pp. 243–272, Dec. 2008.
[34] A.-I. Popa, M. Zanfir, and C. Sminchisescu, "Deep multitask architecture for integrated 2D and 3D human sensing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017.
[35] R. Ranjan, V. M. Patel, and R. Chellappa, "HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.
[36] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1202–1213, Mar. 2018.
[37] A.-A. Liu, Y.-T. Su, W.-Z. Nie, and M. Kankanhalli, "Hierarchical clustering multi-task learning for joint human action grouping and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 102–114, Jan. 2017.
[38] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4570–4579.
[39] P. Dou, S. K. Shah, and I. A. Kakadiaris, "End-to-end 3D face reconstruction with deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017.
[40] M. Tatarchenko, S. R. Richter, R. Ranftl, Z. Li, V. Koltun, and T. Brox, "What do single-view 3D reconstruction networks learn?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3405–3414.
[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980
[42] F. Bernardini, J. Mittleman, H. Rushmeier, C. Silva, and G. Taubin, "The ball-pivoting algorithm for surface reconstruction," *IEEE Trans. Visual. Comput. Graph.*, vol. 5, no. 4, pp. 349–359, Oct. 1999.

FEI HU is currently pursuing the Ph.D. degree with the Key Laboratory of Media Audio and Video, Communication University of China, Ministry of Education, Beijing, China.

He has authored conference publication and coauthored another conference publication. He is currently working with Prof. Zhang on artificial intelligence and virtual reality.

LONG YE received the B.Eng. degree in electronic engineering from Shandong University, Jinan, China, in 2003, and the M.Eng. and Dr.Eng. degrees from the Communication University of China, Beijing, China, in 2006 and 2012, respectively.

He is currently with the Key Laboratory of Media Audio and Video, Communication University of China, Ministry of Education. Prior to that, he was a Visiting Scholar with Ryerson University, Toronto, Canada. His research interests include computer vision, image compression, and virtual reality.

WEI ZHONG received the B.S. degree in electronic engineering and the Ph.D. degree in circuits and systems from Xidian University, in 2004 and 2010, respectively.

She joined the Key Laboratory of Media Audio and Video, Ministry of Education, Communication University of China, in 2011, where she has been an Associate Professor, since 2015. Her research interests focus on the theory and design of multirate filter banks, image processing, and video affective content analysis.

LI FANG received the M.Eng. and Dr.Eng. degrees from the Communication University of China, Beijing, China, in 2006 and 2012, respectively.

He is currently with the Key Laboratory of Media Audio and Video, Communication University of China, Ministry of Education. His research interests include computer vision and virtual reality.

QIN ZHANG received the Ph.D. degree in engineering from The University of British Columbia, Vancouver, BC, Canada, in 1990.

He worked as a Research and Development Scientist with the EE Department, The University of British Columbia, from 1990 to 1995. In 2004, he has served as the Dean of the TCL Industrial Research Institute. He is currently with the Key Laboratory of Media Audio and Video, Communication University of China, Ministry of Education. His research interests include computer vision, image compression, and virtual reality.