IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 28, 2019, accepted November 15, 2019, date of publication November 22, 2019,
date of current version December 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2955288

3D-FHNet: Three-Dimensional Fusion Hierarchical Reconstruction Method for Any Number of Views

QIANG LU[1,2,3], YIYANG LU[2], MINGJIE XIAO[2], XIAOHUI YUAN[4], (Senior Member, IEEE), AND WEI JIA[1,2]

[1]Key Laboratory of Knowledge Engineering With Big Data, Ministry of Education, Hefei University of Technology, Hefei 230009, China
[2]School of Computer and Information, Hefei University of Technology, Hefei 230009, China
[3]Anhui Province Key Laboratory of Industry Safety and Emergency Technology, Hefei University of Technology, Hefei 230009, China
[4]Department of Science and Engineering, University of North Texas, Denton, TX 76203, USA

Corresponding author: Qiang Lu (luqiang@hfut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61972130 and Grant 61673157, in part by the Fundamental Research Funds for Central Universities under Grant PA2018GDQT0014, in part by the Key Research and Development Program in Anhui Province under Grant 1804a09020036, in part by the Key Laboratory of Agricultural Electronic Commerce, Ministry of Agriculture of China, under Grant AEC2018003, and in part by the Key Project of Transformation and Industrialization of Scientific and Technological Achievements of Intelligent Manufacturing Technology Research Institute, Hefei University of Technology, under Grant IMICZ2017010.

ABSTRACT The research field of reconstructing 3D models from 2D images is becoming more and more important. Existing methods typically perform single-view reconstruction or multi-view reconstruction utilizing the properties of recurrent neural networks. Due to the self-occlusion of the model and the special nature of the recurrent neural network, these methods have some problems. We propose a novel three-dimensional fusion hierarchical reconstruction method that utilizes a multi-view feature combination method and a hierarchical prediction strategy to unify the single view and any number of multiple views 3D reconstructions. Experiments show that our method can effectively combine features between different views and obtain better reconstruction results than the baseline, especially in the thin parts of the object. Our source code is available at https://github.com/VIM-Lab/3D-FHNet.

INDEX TERMS 3D reconstruction, multi-views reconstruction, 3D volume, feature combination, hierarchical prediction.

I. INTRODUCTION

With the continuous upgrading of application requirements such as robot grasping objects and 3D printing, it becomes a requirement to automatically reconstruct 3D models from 2d images. Moreover, we can get more information from the three-dimensional model than from the two-dimensional image. Therefore, this research field is becoming more and more important.

The emergence of some large 3D model libraries, such as ShapeNet [1], PASCAL 3D+ [2], and ObjectNet3D [3], has promoted the development of this research field. In recent years, with the rapid development of deep learning, learning-based methods have become the mainstream of 3D reconstruction. These methods usually accept two-dimensional images as input to obtain 3D reconstruction results in voxel [4], mesh [5] and point cloud [6] formats.

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

However, due to the self-occlusion of the model, a single image can correspond to a variety of possible 3D models [7]. Therefore, when the information contained in the single image is limited, it is an impossible task to infer the accurate 3D model. In addition, the result of 3D reconstruction using a single view is also unstable. To solve this problem, it is natural and effective to utilize multiple images of the same object from different perspectives to reconstruct the 3D model.

When we map multiple images of the same object to a 3D model, a new problem arises: how do we combine the information contained in multiple images? The vast majority of current methods [8], [9] use the features of each image as an input to the LSTM or its variants, using LSTM's memory capabilities to combine the information contained in multiple images by constantly updating the hidden state. These methods subtly utilize the characteristics of LSTM to combine the information of multiple images, and as the number of views increases, the accuracy of reconstruction results is improved.

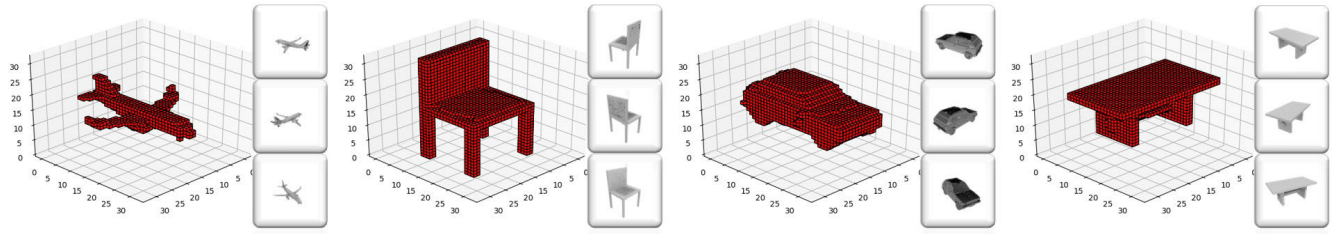


FIGURE 1. Some examples of reconstruction by our 3D-FHNet. The result of the reconstruction is on the left, and the input views are on the right.

However, there are some problems in using the features of each image as an input to the LSTM. As we know, the LSTM structure is time-series. When we use multiple images as input to different time steps of the LSTM, we distinguish the status of each image. Due to the structure of the LSTM, the input of different time steps will have different effects on the final reconstruction result. That is to say, the result of the reconstruction will depend on the order of the input images. If the order of the input images changes, the final reconstruction result is changed accordingly, which is obviously not our intention. When visualizing the reconstruction results, for thin parts of objects such as chair legs, existing methods readily achieve the result of missing components during reconstruction.

Therefore, we propose a novel multi-view feature combination method. This feature combination method treats each input image equally and the reconstruction result will not change due to the change of input image order. Furthermore, the feature combination method can receive any number of images as input and obtain a three-dimensional reconstruction result, and as the number of input images increases, the reconstruction result is improved. In addition, we propose a hierarchical prediction strategy, which can effectively improve the reconstruction results for the thin parts of the object.

Our experiments show that our model can achieve better reconstruction results than the state-of-the-art method, and the reconstruction results will become more and more accurate as of the number of views increases. The main contributions of this paper are as follows:

- We propose a three-dimensional fusion hierarchical reconstruction method called 3D-FHNet, which unifies a single view and any number of multiple views reconstructions and can get accurate reconstruction results.
- We propose a novel multi-view feature combination method that allows our model to continuously improve reconstruction performance as the number of views increases.
- We utilize a hierarchical prediction strategy that allowed the network to reconstruct the thin parts of the object more accurately.
- Our experiments show that our method can achieve better reconstruction results than the state-of-the-art method.

II. RELATED WORK

Classic 3D reconstruction methods, based on the Structure-from-Motion technology [10]–[13] are usually limited to the illumination condition, surface textures and dense views. The goal of these multi-view instance reconstruction methods is to infer the 3D structure of a particular scene/object given a large number of views of the same instance. Contrary to these methods, benefited from prior knowledge, our method can reconstruct credible results with a small number of images or even one image without the assumptions on the object reflection and surface textures. This is something that these classic 3D reconstruction techniques cannot do.

With the emergence of large-scale shape sets [1]–[3], [14], especially the success of CNNs, data-driven methods have become the preferred method to predict 3D shapes. Regarding object reconstruction as a predictive and generative issue from a single image, learning-based methods typically utilize a CNN-based encoder and decoder to predict 3D volumes, meshes, or point sets. Girdhar *et al.* [15] combined an autoencoder and a convolutional network to learn an embedding space with 2D images and 3D shapes. Dai *et al.* [16] completed partial 3D shapes through a combination of volumetric deep neural networks and 3D shape synthesis. Wu *et al.* [17] generated 3D objects from a probabilistic space leveraging volumetric convolutional networks and generative adversarial nets. Smith and Meger [18] extended previous work by employing the Wasserstein distance normalized with gradient penalization as a training objective. Kar *et al.* [19] proposed deformable 3D shape models to recovering high frequency shape details. Zhu *et al.* [4] designed architectures of pose-aware shape reconstruction which reproject the predicted shape back on to the image using the predicted pose. Fan *et al.* [6] designed a conditional shape sampler, capable of predicting plausible 3D point clouds from an input image. Wang *et al.* [20] represented 3D mesh in a graph-based convolutional neural network and produced correct geometry by progressively deforming an ellipsoid, leveraging perceptual features extracted from the input image. Mandikal *et al.* [21] learned mapping from the 2D image to the corresponding learned to embed by learning a probabilistic latent space with a view-specific "diversity loss". Yan *et al.* [22] investigated the task of single-view 3D object reconstruction from a learning agent's perspective. Tulsiani *et al.* [23] studied the notion of consistency between a 3D shape and a 2D observation and proposed a differentiable formulation, which allows

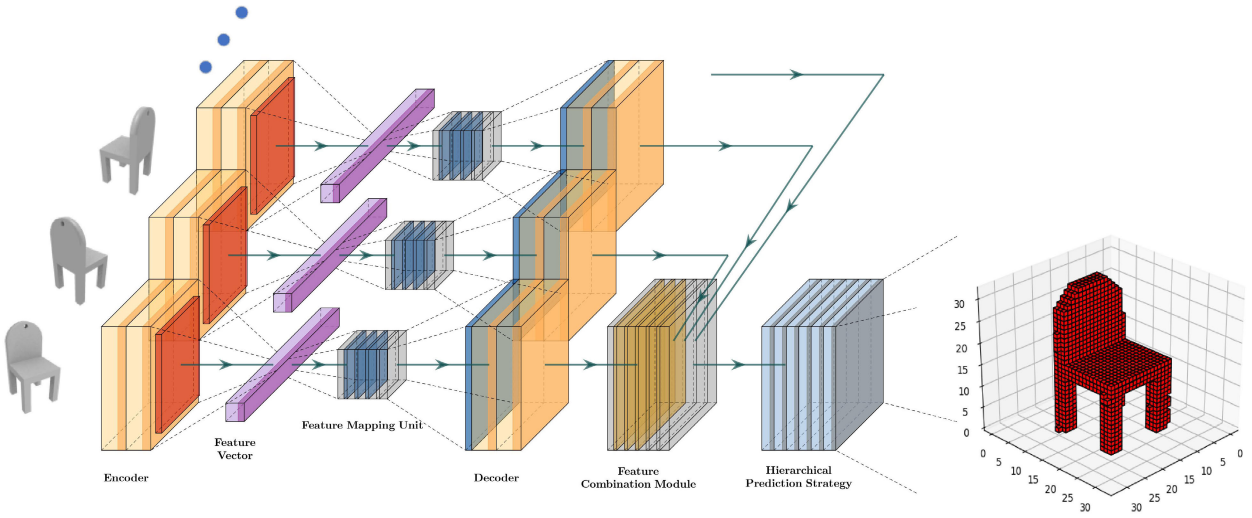


FIGURE 2. We propose a unified framework for performing single view or any number of multiple views reconstructions. In our model, the two-dimensional encoder extracts the image features, mapping them to the three-dimensional features through the feature mapping unit, and decodes the predicted voxel occupancy probability through the three-dimensional decoder. The features of the multiple images are combined by the feature combination module, and finally, the hierarchical prediction strategy is utilized to predict the three-dimensional volume of the object.

computing gradients of the 3D shape given an observation from an arbitrary view. Kato *et al.* [24] proposed an approximate gradient for rasterization that enables the integration of rendering into neural networks. Yang *et al.* [25] reconstructed the complete 3D structure of a given object from a single arbitrary depth view using generative adversarial networks. Kurenkov *et al.* [26] introduced a new differentiable layer for 3D data deformation and used it to learn a model for 3D reconstruction-through-deformation. A crucial assumption in the above-mentioned models, however, is that the input images contain most information of a 3D object. As a result, these models fail to make a reasonable prediction when the observation has severe self-occlusion as they lack the information from other views. Sun *et al.* [27] proposed an end-to-end efficient generation network to reconstruct 3D model from a single image. Reference [28] proposed a self-supervised network to generate 3D point clouds from a single RGB image.

An effective solution is to utilize more views to make up the information. Tulsiani *et al.* [29] allows leveraging multi-view observations from unknown poses as a supervisory signal during training. Kar *et al.* [30] leverage the underlying 3D geometry of the problem through feature projection and unprojection along viewing rays. Choy *et al.* [8] utilize a recurrent neural network to learn a mapping from images of objects to their underlying 3D shapes from a large collection of synthetic data. Yang *et al.* [9] learn a guided information acquisition model and to aggregate information from a sequence of images for reconstruction. Soltani *et al.* [31] learnt a generative model over multi-view depth maps or their corresponding silhouettes, and used a deterministic rendering function to produce 3D shapes from these images. Wiles and Zisserman [32] introduced a deep-learning architecture and loss function, which handle multiple views in an order-agnostic manner. Gwak *et al.* [33] explored inexpensive 2D supervision as an alternative for expensive 3D CAD annotation, used foreground masks as weak supervision through a ray trace pooling layer that enables perspective projection and backpropagation.

Closest to our work is the work of Choy *et al.* [8] which takes in one or more images of an object instance from arbitrary viewpoints, learns a mapping from images of objects to their underlying 3D shapes from a large collection of synthetic data and outputs a reconstruction of the object in the form of a 3D occupancy grid. We utilize a multi-view feature combination method to combine the features of multiple views and a hierarchical prediction strategy to obtain the three-dimensional reconstruction results.

III. METHODOLOGY

In this chapter, we will introduce our model in detail. In section 3.1, an overview of our approach is provided. In Section 3.2, our single-view reconstruction network architecture is described in detail. In Section 3.3, our multi-view feature combination module is described. In Section 3.4, the hierarchical prediction strategy utilized by our model is described.

A. OVERVIEW

We developed a unified framework to perform both single view and any number of multiple views reconstructions. We represent the three-dimensional model as a voxel form, and the voxel is characterized by a value of zero or one for each voxel grid. Benefiting from this feature of voxel representation, we consider the prediction of the voxel grid at each location in the reconstruction task as a binary classification problem.

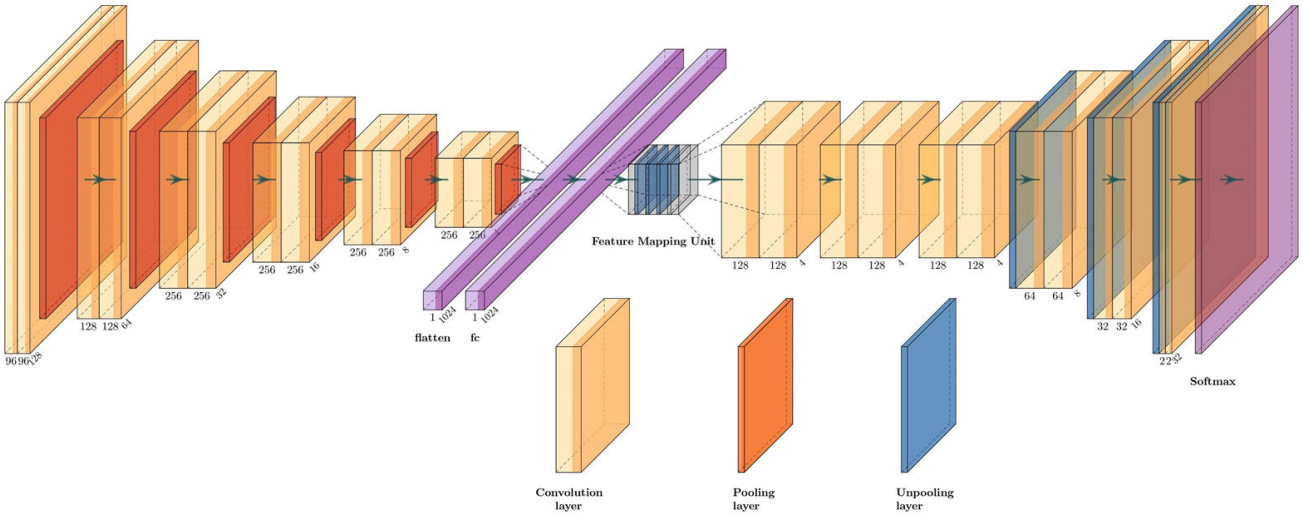


FIGURE 3. Illustration of network architecture. Our entire single view reconstruction network consists of three components: a two-dimensional encoder, a feature mapping unit and a three-dimensional decoder. Taking the RGB image as input, the network predicts the predicted probability O_t of the three-dimensional volume.

Figure 2 illustrates our model. In the training stage, we can feed any number of image data of the same object to our model. These images are respectively encoded by a residual two-dimensional encoder and then flattened into a feature vector F_i. Each F_i is then sent to the feature mapping unit to convert the two-dimensional information into three-dimensional information V_i. This three-dimensional information is then decoded by a residual 3D decoder, resulting in a predicted voxel occupying O_t. As the number of views increases, each predicted voxel occupancy O_t is combined by a feature combination module, and finally the final predicted voxel occupancy O is obtained. Under the guidance of cross-entropy loss function, the parameters in the model are optimized gradually. In the test stage, we feed the images of the test set to the trained model to obtain the predicted voxel occupation O, and then utilize the hierarchical prediction strategy to obtain 0-1 voxel occupation, where zero represents no occupation and 1 represents occupation. Finally, the accuracy is calculated by comparing the predicted 0-1 voxel occupation with the ground truth voxel occupation.

B. SINGLE VIEW RECONSTRUCTION NETWORK ARCHITECTURE

We utilize the encoder-decoder architecture to generate 3D models. We plot one step of data flow in Figure 3. Next, we discuss the detailed architecture.

1) ENCODER: TWO-DIMENSIONAL RESIDUAL CONVOLUTIONAL NEURAL NETWORK

This network is utilized to extract features from the input image. In our implementation, the network was utilized to extract features from images with a resolution of 128×128. For each input image, we make it pass through six residual convolutional encoder blocks successively. Each residual convolutional encoder block consists of three convolution operations and one pooling operation. For each residual convolutional encoder block, we pass the data through two paths at the same time, one of which contains two convolutions and the other is a 1×1 convolution, all convolution operations are followed by a relu activation function. The data of the two paths are concatenated, passed through a max-pooling and outputted to the next operating unit. The convolution kernel size in the first residual convolutional encoder block is set to 7×7, and the convolution kernel size in the remaining five residual convolutional encoder blocks is set to 3×3. After passing through six residual convolutional encoder blocks, we flatten the extracted features into feature vectors.

2) FEATURE MAPPING UNIT

The feature mapping unit is utilized to map the feature vector extracted by the encoder to the three-dimensional feature. For each feature vector of the input image, we pass it through a fully connected layer and then send it to the feature mapping unit. We set up a W matrix, and a B matrix, for mapping the two-dimensional feature vector F to the three-dimensional feature V. In our implementation, we set the dimension of the output of the fully connected layer to 1×1024, the dimension of the W matrix to $4 \times 4 \times 4 \times 1024 \times 128$, and the dimension of the B matrix to $4 \times 4 \times 4 \times 128$. The two-dimensional feature vector F is mapped to the three-dimensional feature V by

$$V_{i,j,k} = W_{i,j,k} \times F + B_{i,j,k} \tag{1}$$

where $V_{i,j,k}$ represents the three-dimensional feature of the corresponding position of the i, j, k coordinates in the three-dimensional feature V. $W_{i,j,k}$ represents a weight matrix in the feature mapping unit for mapping to the three-dimensional feature $V_{i,j,k}$ of the corresponding position of the i, j, k coordinates. $B_{i,j,k}$ represents a bias matrix in the feature

mapping unit when mapped to the three-dimensional feature $V_{i,j,k}$ corresponding to the position of i, j, k coordinates.

3) DECODER: 3D RESIDUAL DECONVOLUTIONAL NEURAL NETWORK

This network is utilized to decode three-dimensional features into three-dimensional volumes. For each input image, the network acquires the three-dimensional feature V_i outputted by the feature mapping unit as an input, sequentially passes through six three-dimensional residual deconvolutional decoder blocks, and normalizes by softmax to obtain a three-dimensional volume prediction probability O_t. Each three-dimensional residual deconvolutional decoder blocks consists of three deconvolution operations and one unpooling operation. For each residual deconvolutional decoder block, we pass the data through two paths at the same time, one of which contains two three-dimensional convolutions, the other is a $1 \times 1 \times 1$ convolution, all convolution operations are followed by a relu activation function. The data of the two paths are concatenated, passed through an unpooling, and outputted to the next operating unit. The convolution kernel size of the three-dimensional convolution in each three-dimensional residual deconvolutional decoder block is set to $3 \times 3 \times 3$.

Since we utilize the characteristics of 3D voxel representation to transform 3D reconstruction into multiple binary classification tasks, we can define the loss function as the cross-entropy loss function commonly utilized in classification tasks.

More formally,

$$\text{Loss} = -\sum_{i=1}^{32}\sum_{j=1}^{32}\sum_{k=1}^{32}\left[GT_{i,j,k} \times \ln O_{i,j,k} + \left(1 - GT_{i,j,k}\right) \times \ln\left(1 - O_{i,j,k}\right)\right] \quad (2)$$

where $GT_{i,j,k}$ represents the value of the voxel grid corresponding to the coordinate position of i, j, k in the ground truth. $O_{i,j,k}$ represents the predicted probability of the voxel grid corresponding to the coordinate position of i, j, k in the final predicted voxel occupancy probability.

If our model is performing single-view 3D reconstruction, then the prediction probability O_t is the final predicted voxel occupancy probability O, and then we will utilize our hierarchical prediction strategy for prediction; if our model is performing multi-view 3D reconstruction, then The predicted probability O_t obtained by each view will be sent to the multi-view feature combination module to obtain the final predicted voxel occupancy probability O, and then utilize our hierarchical prediction strategy for prediction.

C. MULTI-VIEW FEATURE COMBINATION MODULE

Multi-view feature combination module is used to combine features of multiple input images. When we see an object, we can know the general shape of the object. We have a good grasp of the shape of the object directly exposed to us. For the part of the object that is not visible due to self-occlusion, we can probably guess its shape according to our life experience, but it is not very sure. When we walk around the object, we can know exactly what the object is like. Inspired by this phenomenon, when our model sees a picture, we can get the current predicted voxel occupancy probability O_t. For this O_t, the part of the picture that can be directly observed in this picture will get a more certain probability, while the part occluded by the object itself, is not visible in this view, our model will try to guess the occupancy probability of each voxel grid based on prior knowledge. That is to say, the part of the reconstructed object that can be directly observed in an image, the predicted occupancy probability will be closer to 1, and the part that is not directly observable due to self-occlusion, the predicted occupancy probability will be a more conservative prediction based on prior knowledge.

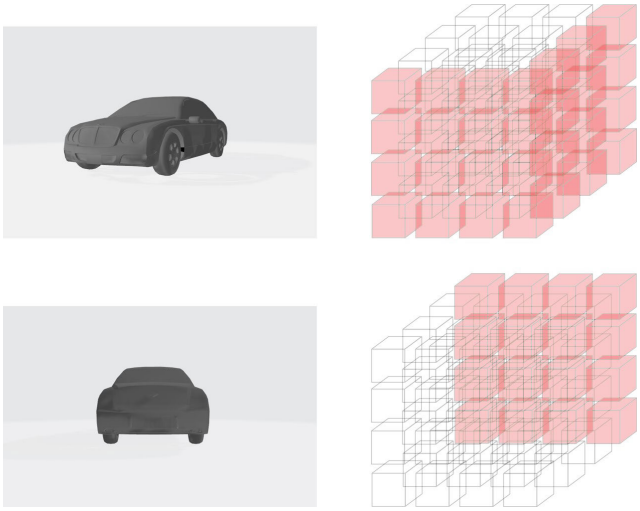


FIGURE 4. For an input image, the model is biased to determine the shape of the visible portion of the reconstructed object, while for parts that are not visible due to self-occlusion, the model will make a conservative estimate of its shape. For example, when the image of the upper left part is input, the model is biased to determine the shape of the front and left of the car. That is, the red part of the upper right picture; when the image of the lower-left part is input, the model will be biased to determine the shape of the back of the car. That is, the red part of the lower right picture.

For each input image, we obtain the predicted voxel occupation probability O_t, whose predicted voxel occupation probability will be biased to the visible part of the object, as shown in figure 4. As the number of input views increases, the visible portion of the object increases as the camera position of each input image is different. When our model gets information from more and more images, each input image can make our model convinced that the voxel grids of some coordinates are occupied, while the occupation of the voxel grids of other coordinates is temporarily undetermined. As the number of views increases, more and more voxel grids can be determined, and the reconstruction performance of our models is constantly improving.

More formally, for the final predicted voxel occupancy probability O, we let

$$O_{i,j,k} = \max_{1 \le t \le n} O^{(t)}_{i,j,k} \quad (3)$$

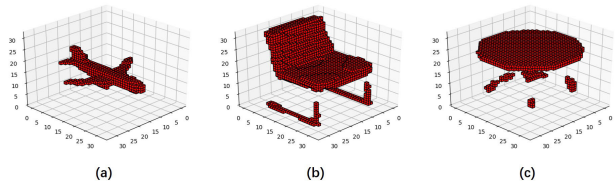


FIGURE 5. Some reconstruction examples that failed in the detail sections. (a) failed to reconstruct the landing gear of the aircraft (b) failed to reconstruct the complete chair legs (c) failed to reconstruct the complete table legs.

where $O_{i,j,k}$ represents the predicted probability of the voxel grid corresponding to the coordinate position of i, j, k in the final predicted voxel occupancy probability. $O^{(t)}_{i,j,k}$ represents the predicted probability of the voxel grid corresponding to the coordinate position of i, j, k in the predicted voxel occupancy probability obtained from the i-th image. $\max_{1\leq t\leq n} O^{(t)}_{i,j,k}$ represents the maximum value among the n $O^{(t)}_{i,j,k}$ obtained in a total of n input images.

We will verify the performance of our multi-view feature combination module in Section 4.4.

D. HIERARCHICAL PREDICTION STRATEGY

After obtaining the final three-dimensional volume prediction probability, the general method sets a threshold, such as 0.5, predicts the voxel grid with the predicted probability greater than or equal to the threshold as occupied, and predicts the voxel grid with the predicted probability less than the threshold as unoccupied. When analyzing the visualization results of the reconstruction, we found that in the thin part of the object, it is easy to get the failed reconstruction result. Benefiting from our multi-view feature combination module, we have improved our performance on this issue. Utilizing our proposed hierarchical prediction strategy, we can further improve the reconstruction results of our model in the thin part of the object.

As shown in Figure 5, the details of the small parts of the object such as the legs of the chair, the legs of the table, and aircraft landing gear are the parts of the model that are most likely to fail. The failure reconstruction results mostly in the absence of details. For example, the reconstructed chair has no chair legs, only the backrest and seat of the chair remained, or the thick table and table legs get a thinner reconstruction result. In response to this issue, we propose a hierarchical prediction strategy. After obtaining the final three-dimensional volume prediction probability O, we determine whether these voxel grids are occupied from the outer voxel grid to the inner voxel grid layer by layer. When determining whether a voxel grid is occupied, we look at the occupancy of the voxel grid on the outer layer and dynamically adjust the threshold of the voxel grid according to the occupancy of the outer voxel grid. If the outer voxel grid is occupied by a small number, a smaller threshold is used; otherwise, a larger threshold is used. That is, if a voxel is located on the surface of the object, our model is more likely to predict it as occupied, and as the prediction enters the interior of the object, the prediction will gradually become unbiased.

We evaluate the performance of our hierarchical prediction strategy in Section 4.5.

IV. EVALUATION

In this section, we discuss the following three questions: (1) Can our network generate more accurate reconstruction results? (section 4.3) (2) Can our network improve the accuracy of reconstruction results as the number of views increases? (section 4.4) (3) Can our hierarchical prediction strategy optimize reconstruction results? (section 4.5)

A. DATASET

The ShapeNet dataset is a collection of 3D CAD models organized according to the WordNet hierarchy. ShapeNetCore is a subset of the full ShapeNet dataset with single clean 3D models and manually verified category and alignment annotations. It covers 55 common object categories with about 51,300 unique 3D models. Since most of the 55 common object categories contained too few 3D models, we selected 13 categories with more than 1,000 3D models in the ShapeNetCore dataset. The 13 categories are plane, car, chair, sofa, table, bench, cabinet, monitor, lamp, speaker, rifle, telephone, and vessel, which contain a total of 43,783 3D models. For each 3D model, we rendered 12 images of different angles with a resolution of 128^2 and generated a ground truth voxel occupation with a resolution of 32^3. We refer to this dataset as the ShapeNet dataset throughout the evaluation section. We divided ShapeNet dataset into a training set, test set, and validation set according to the proportion of 80%, 16%, and 4%, and refer to these three datasets as the ShapeNet training set, ShapeNet test set and ShapeNet validation set throughout the evaluation section.

B. BASELINE

The state-of-the-art method that relevant to our method is 3D-R2N2 [8]. We compare our method with 3D-R2N2, which performs both single and multi-view 3D reconstruction using a 3D recurrent network, combining features by using features of multiple views as multiple inputs to the LSTM. For a fair comparison, we trained both 3D-R2N2 and our 3D-FHNet on the same ShapeNet training set for the same iterations and ensured that both models have converged. In the test stage, the two trained models were tested on ShapeNet test set and evaluated the qualitative results and the quantitative results separately. We also compared our method with single view reconstruction methods that generate 32^3 voxel reconstruction results. Our comparison targets are PTN(Perspective Transformer Network) [22] and OGN(Octree Generating Networks) [34].

C. EVALUATION ON RECONSTRUCTION PERFORMANCE

We use the voxel IoU (intersection-over-union) as an indicator to quantitatively assess the performance of reconstruction. The voxel IoU is a widely used indicator to measure the final

TABLE 1. Comparison of our method and 3D-R2N2 per-category reconstruction performance on different numbers of views. Except for the reconstruction of the single view speaker class, our method has much better reconstruction performance than 3D-R2N2.

views	1		3		6		12	
method	R2N2	FHNet	R2N2	FHNet	R2N2	FHNet	R2N2	FHNet
plane	0.654	**0.697**	0.672	**0.700**	0.682	**0.709**	0.677	**0.734**
car	0.866	**0.871**	0.878	**0.884**	0.884	**0.885**	0.874	**0.891**
chair	0.549	**0.589**	0.576	**0.625**	0.600	**0.628**	0.583	**0.649**
sofa	0.733	**0.750**	0.768	**0.782**	0.778	**0.803**	0.747	**0.794**
table	0.611	**0.652**	0.608	**0.649**	0.632	**0.657**	0.639	**0.668**
bench	0.549	**0.570**	0.539	**0.603**	0.554	**0.585**	0.512	**0.646**
cabinet	0.783	**0.797**	0.814	**0.845**	0.801	**0.862**	0.806	**0.848**
monitor	0.598	**0.625**	0.626	**0.680**	0.655	**0.683**	0.629	**0.657**
lamp	0.412	**0.480**	0.461	**0.511**	0.426	**0.516**	0.442	**0.531**
speaker	**0.752**	0.748	0.749	**0.775**	0.760	**0.776**	0.746	**0.803**
rifle	0.645	**0.668**	0.619	**0.684**	0.658	**0.684**	0.634	**0.709**
telephone	0.784	**0.807**	0.786	**0.841**	0.843	**0.872**	0.818	**0.853**
vessel	0.618	**0.631**	0.629	**0.666**	0.623	**0.658**	0.613	**0.684**
average	0.662	**0.688**	0.677	**0.713**	0.688	**0.715**	0.680	**0.728**

predicted voxel occupancy. The value is the number of voxel grids in the intersection of all voxel grids predicted to be occupied and all voxel grids occupied by the ground truth values, divide by the number of voxel grids in the union of all voxel grids predicted to be occupied and all voxel grids occupied by the ground truth values.

More formally,

$$IoU = \frac{Prediction \cap GroundTruth}{Prediction \cup GroundTruth} \tag{4}$$

where Prediction refers to the final predicted 0-1 voxel occupancy, and GroundTruth refers to the ground truth 0-1 voxel occupancy.

The voxel IoU punishes the wrong result in two ways. If the model predicts a voxel grid with a ground truth value of 0 as 1, it will make the union in the denominator larger; Conversely, if the model predicts a voxel grid with a ground truth of 1 as 0, the intersection in the numerator will be smaller. Therefore, the range of voxel IoU is [0, 1], and the larger the value, the higher the accuracy of the model.

1) PER-CATEGORY RESULTS

We evaluated the reconstruction performance of 13 categories on the ShapeNet test set. The evaluation results of IoU are shown in Table 1.

It can be seen that in the single view reconstruction, our method is better than 3D-R2N2. In the reconstruction of the speaker class, the reconstruction performance of our method is slightly weaker than that of 3D-R2N2. In other categories, the reconstruction performance of our method is much better than that of 3D-R2N2. This can be attributed to our feature mapping unit and hierarchical prediction strategy. Relative to most models that can perform multi-view reconstruction, we use feature mapping unit to replace LSTM-based structures. Compared to LSTM-based models, our feature mapping unit can retain more information when mapping 2D features to 3D features. Moreover, we use a hierarchical prediction strategy to partially overcome the lack of information due to view scarcity, further enhancing our single view reconstruction performance.

In multi-view reconstruction, our method performs much better than 3D-R2N2 in all categories, and it has more advantages than single-view reconstruction. The average reconstruction performance of our method has more than 7% improvement over 3D-R2N2 when providing all views of the object to the model. This can be attributed to our multi-view feature combination module. As the number of views increases, our model can better combine the information of different views to improve the reconstruction performance compared to the LSTM-based method.

We also calculated the F-scores of our method, which are comprehensive consideration of Precision and Recall.

$$F - Score = \left(1 + \beta^2\right) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \tag{5}$$

In the case of inputting 12 views, the average F1-Score of the results obtained by our method is 0.828, and the average F2-Score is 0.846. In comparison, the average F1-Score of the results obtained by 3D-R2N2 is 0.789, and the average F2-Score is 0.783.

2) QUALITATIVE RESULTS

The example of reconstruction results shown in figure 6 qualitatively shows that our model can reconstruct better three-dimensional models from two-dimensional images. In Figure 6, the columns from left to right are the input images, the ground truth voxel occupancy, the reconstruction result of our method, and the reconstruction result of 3D-R2N2. It can be seen that our method performs better than 3D-R2N2 in the reconstruction of each category. In the reconstruction of the aircraft, the results of our method are better in details such as the landing gear and the curvature of the wings. When reconstructing the bench, the results of our method are better in details such as the handrails and legs of the bench. When reconstructing the cabinet, the reconstruction of our method performed better in the details of the hollowing out of the cabinet and the top of the cabinet. In the reconstruction of the chair, the reconstruction results of our method are better in details such as the legs of the chair. In the reconstruction of the lamp, the reconstruction results

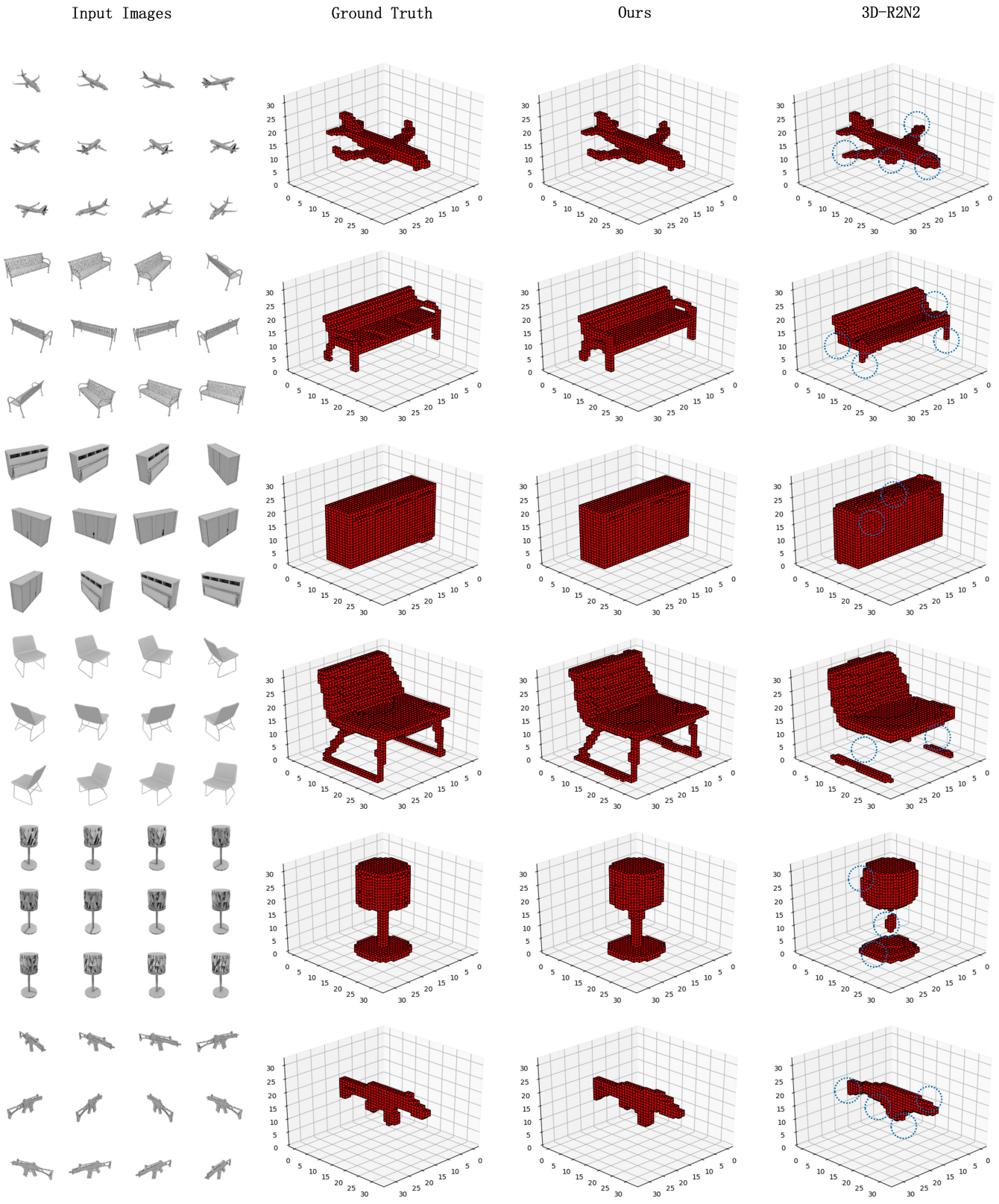


FIGURE 6. Qualitative comparison of our method with 3D-R2N2. The columns from left to right are the input images, the ground truth, the reconstruction result of our method, and the reconstruction result of 3D-R2N2.

of our method are better in the lampshade, base, pole and other details. In the reconstruction of the rifle, our method performed better in the details of the grip, butt, and head of the rifle.

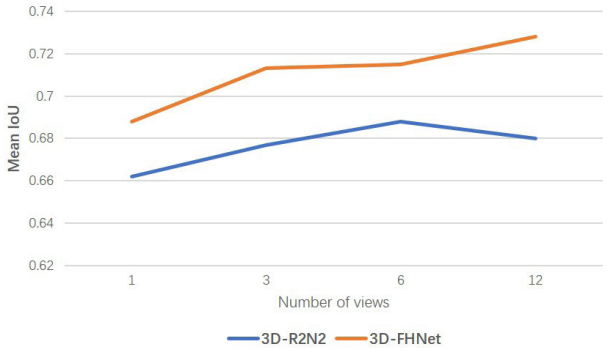


FIGURE 7. As the number of views increases, the reconstruction performance of our network continues to increase and is always better than the performance of 3D-R2N2. In contrast, when the number of views is small, the reconstruction performance of 3D-R2N2 increases with the number of views. However, as the number of views continues to increase, its performance has not continued to increase, but has declined somewhat.

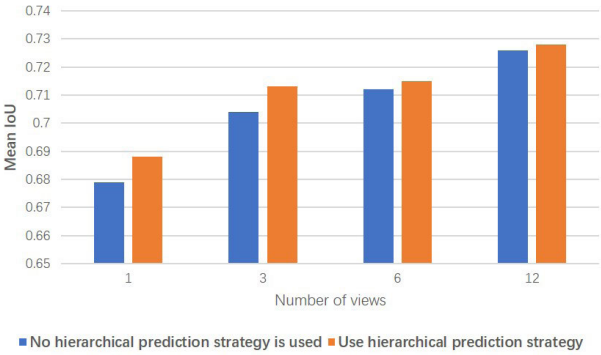


FIGURE 8. Our hierarchical prediction strategy can effectively improve reconstruction performance, especially when the number of input views is small and many details of objects cannot be determined.

D. EVALUATION ON MULTI-VIEW FEATURES COMBINING

We compared our method with 3D-R2N2 in multi-view feature combination performance. Most of the previous methods, including 3D-R2N2, use the features of different input images as inputs on different time steps of the LSTM, utilizing the memory function of LSTM, combining the features of different views. We believe that these LSTM-based methods lose some of the information when performing a feature combination because the input images are not time-series. In contrast, our method utilizes our multi-view feature combination module to combine features of different views.

We reconstructed the single view, 3 views, 6 views, and 12 views, respectively, and compared the reconstruction performance of the two methods. The results are shown in Figure 7. It can be seen that when the number of views is small, the reconstruction performance of our method and 3D-R2N2 increases as the number of views increases; When the number of views continues to increase, 3D-R2N2 fails in combining the features of more views, and its reconstruction performance does not continue to improve, but decreases. In contrast, our method continues to improve as the number of views continues to increase. This can be attributed to the fact that our multi-view feature combination method has better feature combination ability than LSTM. Using the features of the different views as inputs to different time steps of the LSTM can combine the features of the different views and get the final prediction. However, due to the forgetting and updating mechanism, the LSTM-based methods rely on the order of input views when combining features of multiple views, and will discard some information of early views when there are too many time steps. In contrast, our multi-view feature combination method does not depend on the order of the input views and does not discard any information from any of the views, which allows our model to achieve better reconstruction results in multi-view reconstruction.

We also compared our approach to single-view reconstruction methods, which were able to generate 32^3 voxel resolution reconstruction results. The reconstruction results obtained by these single-view reconstruction methods are usually better than the single-view reconstruction results of the multi-view reconstruction method. The comparison results are shown in Table 2.

As can be seen from the table, different methods have some advantages in different categories when performing the single-view reconstruction. Overall, OGN performs best. However, as the number of views increases, our method can outperform the single view reconstruction method in most categories. This shows that our method can effectively combine the information of different views to get more accurate prediction results.

E. EVALUATION ON HIERARCHICAL PREDICTION STRATEGY

We evaluated the performance of our hierarchical prediction strategy to see if it could improve the reconstruction performance of our method.

We performed single view, 3 views, 6 views, and 12 views reconstruction, one using hierarchical prediction strategy, and the other without using hierarchical prediction strategy. The reconstruction performance is shown in Figure 8. It can be seen that in the reconstruction of a different number of views, the model using the hierarchical prediction strategy can perform better. When the number of input views is very large, the use of hierarchical prediction strategies has limited improvement in reconstruction performance because the model determines most of the details of reconstructed objects. However, when the number of input views is limited, our hierarchical prediction strategy can effectively improve the reconstruction performance because the reconstruction object has many details that cannot be determined.

To visually demonstrate the effectiveness of our hierarchical prediction strategy, we utilized our model to perform single-view reconstruction, one using our hierarchical prediction strategy and the other without using our hierarchical prediction strategy. We have chosen an example of reconstruction on the same image of the same object, as shown in Figure 9. On the left side of the figure is the ground truth voxel occupancy. In the middle of the figure is the result of single-view reconstruction without using the hierarchical

TABLE 2. Comparison of the reconstruction performance of our method and single-view reconstruction methods. As the number of views increases, the reconstruction performance of our method surpasses the single-view reconstruction methods that can generate 32^3 voxel reconstruction results.

views	1			3	6	12
method	PTN	OGN	FHNet	FHNet	FHNet	FHNet
plane	0.705	**0.748**	0.697	0.700	0.709	0.734
car	0.773	**0.886**	0.871	0.884	0.885	**0.891**
chair	0.540	0.569	**0.589**	0.625	0.628	**0.649**
sofa	0.751	**0.754**	0.750	0.782	**0.803**	0.794
table	0.561	0.638	**0.652**	0.649	0.657	**0.668**
bench	**0.629**	0.627	0.570	0.603	0.585	**0.646**
cabinet	0.778	**0.797**	**0.797**	0.845	**0.862**	0.848
monitor	**0.684**	0.641	0.625	0.680	0.683	0.657
lamp	0.383	0.430	**0.480**	0.511	0.516	**0.531**
speaker	0.666	0.724	**0.748**	0.775	0.776	**0.803**
rifle	0.690	**0.703**	0.668	0.684	0.684	**0.709**
telephone	**0.863**	0.833	0.807	0.841	**0.872**	0.853
vessel	0.646	**0.761**	0.631	0.666	0.658	0.684
average	0.647	**0.700**	0.688	0.713	0.715	**0.728**

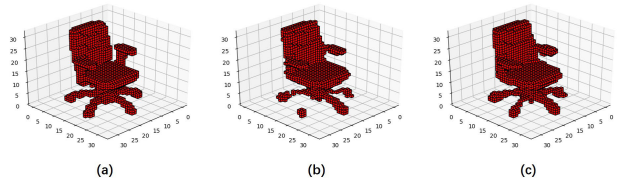


FIGURE 9. A comparison example of using hierarchical prediction strategy and not using hierarchical prediction strategy. (a) is the ground truth voxel occupancy; (b) is the reconstruction result without using the hierarchical prediction strategy; (c) is the reconstruction result using the hierarchical prediction strategy.

prediction strategy. On the right side of the figure is the result of single-view reconstruction using the hierarchical prediction strategy. As can be seen in the figure, in this example, when our model does not use the hierarchical prediction strategy, the single-view reconstruction result is not very good, such as the leg of the office chair is missing. When our model uses the hierarchical prediction strategy for reconstruction, the reconstruction performance is not particularly good due to the single-view reconstruction. However, more details have been reconstructed than the result of the reconstruction without using the hierarchical prediction strategy. Due to the scarcity of views, the information that can be obtained from it is limited. It is impossible to know exactly what its voxel occupancy is in the part of the reconstructed object that is not directly exposed to the view, so our model can only make a conservative prediction. The hierarchical prediction strategy we use dynamically adjusts the thresholds at the time of prediction so that our model can reconstruct more details of the reconstructed objects.

V. CONCLUSION

In this paper, we propose 3D-FHNet, which is a 3D fusion Hierarchical reconstruction method that can perform 3D reconstructions of any number of views. The model unifies single-view and multi-view 3D reconstruction and proposes a multi-view feature combination method to overcome the shortcomings of multi-view feature combination method based on RNN and its variants. In the reconstruction of the thin parts that failed by other methods, the hierarchical prediction strategy is utilized to improve the accuracy. Experiments show that our model can generate more accurate reconstruction results, and as the number of views increases, the accuracy of reconstruction results is further improved. Our model has also achieved better results in the reconstruction of thin parts of the object. In the future, we will try to utilize more efficient data representation to improve the output resolution, overcoming the problem of voxel representation requiring too much memory.

REFERENCES

[1] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "ShapeNet: An information-rich 3D model repository," 2015, *arXiv:1512.03012*. [Online]. Available: https://arxiv.org/abs/1512.03012

[2] Y. Xiang, R. Mottaghi, and S. Savarese, "Beyond PASCAL: A benchmark for 3D object detection in the wild," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2014, pp. 75–82.

[3] 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.* Springer, 2016, pp. 160–176. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-46484-8_10

[4] R. Zhu, H. K. Galoogahi, C. Wang, and S. Lucey, "Rethinking reprojection: Closing the loop for pose-aware shape reconstruction from a single image," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 57–65.

[5] T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry, "A Papier-Mâché approach to learning 3D surface generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 216–224.

[6] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3D object reconstruction from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 605–613.

[7] J. Wu, C. Zhang, X. Zhang, Z. Zhang, W. T. Freeman, and J. B. Tenenbaum, "Learning shape priors for single-view 3D completion and reconstruction," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 646–662.

[8] 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. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 628–644. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-46484-8_38

[9] X. Yang, Y. Wang, Y. Wang, B. Yin, Q. Zhang, X. Wei, and H. Fu, "Active object reconstruction using a guided view planner," 2018, *arXiv:1805.03081*. [Online]. Available: https://arxiv.org/abs/1805.03081

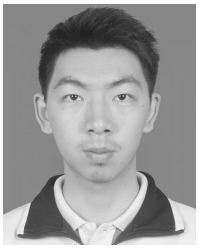
[10] M. Halber and T. Funkhouser, "Fine-to-coarse global registration of RGB-D scans," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1755–1764.

[11] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3D," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 835–846, 2006.

[12] A. Broadhurst, T. W. Drummond, and R. Cipolla, "A probabilistic framework for space carving," in *Proc. 8th IEEE Int. Conf. Comput. Vis. ICCV*, vol. 1, Jul. 2001, pp. 388–393.
[13] X. Chen, Q. Wu, and S. Wang, "Research on 3D reconstruction based on multiple views," in *Proc. 13th Int. Conf. Comput. Sci. Educ. (ICCSE)*, Aug. 2018, pp. 1–5.
[14] X. Sun, J. Wu, X. Zhang, Z. Zhang, C. Zhang, T. Xue, J. B. Tenenbaum, and W. T. Freeman, "Pix3D: Dataset and methods for single-image 3D shape modeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2974–2983.
[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. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-46466-4_29
[16] A. Dai, C. R. Qi, and M. Niessner, "Shape completion using 3D-encoder-predictor CNNs and shape synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5868–5877.
[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] E. Smith and D. Meger, "Improved adversarial systems for 3D object generation and reconstruction," 2017, *arXiv:1707.09557*. [Online]. Available: https://arxiv.org/abs/1707.09557
[19] A. Kar, S. Tulsiani, J. Carreira, and J. Malik, "Category-specific object reconstruction from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1966–1974.
[20] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang, "Pixel2Mesh: Generating 3D mesh models from single RGB images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 52–67.
[21] P. Mandikal, K. L. Navaneet, M. Agarwal, and R. V. Babu, "3D-LMNet: Latent embedding matching for accurate and diverse 3D point cloud reconstruction from a single image," 2018, *arXiv:1807.07796*. [Online]. Available: https://arxiv.org/abs/1807.07796
[22] 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.
[23] S. Tulsiani, T. Zhou, A. A. Efros, and J. Malik, "Multi-view supervision for single-view reconstruction via differentiable ray consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2626–2634.
[24] H. Kato, Y. Ushiku, and T. Harada, "Neural 3D mesh renderer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3907–3916.
[25] B. Yang, H. Wen, S. Wang, R. Clark, A. Markham, and N. Trigoni, "3D object reconstruction from a single depth view with adversarial learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 679–688.
[26] A. Kurenkov, J. Ji, A. Garg, V. Mehta, J. Gwak, C. Choy, and S. Savarese, "Deformnet: Free-form deformation network for 3d shape reconstruction from a single image," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 858–866.
[27] R. Sun, Y. Gao, Z. Fang, A. Wang, and C. Zhong, "SSL-Net: Point-cloud generation network with self-supervised learning," *IEEE Access*, vol. 7, pp. 82206–82217, 2019.
[28] Y. Zhang, Z. Liu, T. Liu, B. Peng, and X. Li, "RealPoint3D: An efficient generation network for 3D object reconstruction from a single image," *IEEE Access*, vol. 7, pp. 57539–57549, 2019.
[29] S. Tulsiani, A. A. Efros, and J. Malik, "Multi-view consistency as supervisory signal for learning shape and pose prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2897–2905.
[30] A. Kar, C. Häne, and J. Malik, "Learning a multi-view stereo machine," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 365–376.
[31] A. A. Soltani, H. Huang, J. Wu, T. D. Kulkarni, and J. B. Tenenbaum, "Synthesizing 3D shapes via modeling multi-view depth maps and silhouettes with deep generative networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1511–1519.
[32] O. Wiles and A. Zisserman, "SilNet: Single- and multi-view reconstruction by learning from silhouettes," 2017, *arXiv:1711.07888*. [Online]. Available: https://arxiv.org/abs/1711.07888
[33] J. Y. Gwak, C. B. Choy, M. Chandraker, A. Garg, and S. Savarese, "Weakly supervised 3D reconstruction with adversarial constraint," in *Proc. Int. Conf. 3D Vis. (3DV)*, Oct. 2017, pp. 263–272.
[34] M. Tatarchenko, A. Dosovitskiy, and T. Brox, "Octree generating networks: Efficient convolutional architectures for high-resolution 3D outputs," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2088–2096.

QIANG LU received the master's and Ph.D. degrees in computer science and information from the Hefei University of Technology. He was a Visiting Scholar with the University of North Texas. He is an Associate Professor with the School of Computer and Information, Hefei University of Technology, where he is a member of the Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education. His primary research interests include visualization, computer graphics, and cooperative computing. He is a member of the CCF.

YIYANG LU was born in China, in 1996. He received the B.S. degree from the Hefei University of Technology (HFUT), Hefei, China, in 2018, where he is currently pursuing the master's degree with the School of Computer and Information. His research interests include 3D reconstruction, computer vision, and machine learning.



MINGJIE XIAO was born in China, in 1995. He received the B.S. degree from the Hefei University of Technology (HFUT), Hefei, China, in 2018, where he is currently pursuing the master's degree with the School of Computer and Information. His research interests include 3D reconstruction, computer vision, and machine learning.

XIAOHUI YUAN (S'01–M'05–SM'16) received the B.S. degree in electrical engineering from the Hefei University of Technology, China, in 1996, and the Ph.D. degree in computer science from Tulane University, in 2004. His research findings have been published in more than 140 peer-reviewed articles. His research interests include computer vision, artificial intelligence, data mining, and machine learning. He is currently an Associate Professor with the University of North Texas. He was a recipient of the Ralph E. Powe Junior Faculty Enhancement Award, in 2008. He serves as the chair of several international conferences. He serves on the editorial board of several international journals. He is the Editor-in-Chief of the *International Journal of Smart Sensor Technologies and Applications*.

WEI JIA received the B.Sc. degree in informatics from Central China Normal University, Wuhan, China, in 1998, the M.Sc. degree in computer science from the Hefei University of Technology, Hefei, China, in 2004, and the Ph.D. degree in pattern recognition and intelligence system from the University of Science and Technology of China, Hefei, China, in 2008. He was a Research Assistant and an Associate Professor with the Hefei Institutes of Physical Science, Chinese Academy of Science, from 2008 to 2016. He is currently an Associate Professor with the School of Computer and Information, Hefei University of Technology. His research interests include computer vision, biometrics, pattern recognition, image processing, and machine learning.