# Single-View and Multi-View 3D Object Reconstruction from Shape Priors

**Xiaobing Zhang, Tengfei Xiao**

*Abstract*—**Deep learning has been widely applied to multi-view 3D reconstruction tasks and has achieved significant progress. The mainstream solutions mainly rely on 2D encoder 3D decoder network architectures to establish mappings between views and object shapes. However, these methods are often limited by image quality and quantity when processing image feature collections, resulting in low-quality 3D shape reconstructions. Humans typically use incomplete or noisy visual cues to retrieve similar 3D shapes from memory and reconstruct the 3D shape of an object. Inspired by this, we propose a new method called RSP3D, which explicitly constructs shape priors to compensate for missing information in images. The shape priors exist in the form of "image-voxel" pairs in a memory network and are used to retrieve accurate 3D shapes that are highly related to the input image. Additionally, we extract information from the retrieved 3D shapes that is useful for object recovery. Experimental results indicate that RSP3D significantly improves the quality of 3D reconstruction.**

***Index Terms***—**3D reconstructions, Multi View, priors, memory network.**

## I. Introduction

3D reconstruction is a cross-cutting issue in the fields of computer vision and computer graphics, and is also the core of many technologies such as computer-aided geometric design, computer animation, medical image processing, digital media, and robotics. As a generative task, transforming a 2D image into a 3D object is undoubtedly a challenging ill-posed inverse problem compared to image restoration and other tasks. According to the number of input images, the task is divided into single-view reconstruction and multi-view reconstruction. This article focuses on deep learning-based multi-view 3D reconstruction algorithms, aiming to reconstruct the shape of 3D objects with voxel representation from multiple images.

At present, most mainstream solutions adopt a basic framework that combines 2D encoders and 3D decoders, and reshapes their advanced features into two-dimensional connections to establish mappings between images and voxels. However, multi-view reconstruction still faces a key problem - how to effectively aggregate features from any number of views.

In our research, there are four types of fusion strategies. After connecting the feature maps from all views, we adjusted the pooling-based method to use a pooling layer to compress the feature maps to a specific size. However, this dimension reduction is too coarse and leads to significant content loss that cannot be avoided. To enable the fusion module to learn, the 3D-R2N2 series used a method based on recurrent neural networks (RNNs). The features from all views are treated as a sequence and processed by a recurrent unit before the decoder. However, this method demonstrates inconsistent predictions for different arrangements. Additionally, due to limited long-term memory, this method is not suitable for numerous views as input. To address these shortcomings, an attention-based fusion method creates a subnetwork to predict the confidence score map for each view and merges features based on this. The AttSets and Pix2Vox series that follow this idea produce stable reconstructors. The former merges features, while the latter mergers voxels directly recovered from each view. Recently, some studies have used transformer architectures for multi-view reconstruction. While leveraging natural advantages, the fusion process is integrated into the encoder stage. They perform well when a large number of views are input, but the reconstruction quality is poor when there are few input images.

We believe that attention-based fusion performs better and more stably compared to other fusion methods, but it still has obvious shortcomings. During the prediction of score maps, the connections of branches rely solely on the softmax layer and there are no learnable parameters, Therefore, it cannot adapt to the global state and only trusts the memory of the net work.To improve this method, we consider integrating shape priors into the attention-based fusion process. Specifically, we first extract image features from given images through a 2D encoder, and then capture the relationship between image features and 3D prototypes with the help of attention mecha nisms. In this way, we can obtain shape priors.Next, we apply clustering algorithms to the shape priors to obtain preliminary representations of objects in 3D space. This representation can help us better understand image features and allocate them reasonably to corresponding 3D prototypes.Finally, we combine the obtained shape priors with attention-based fusion methods to further optimize the matching process between image features and 3D prototypes. In this way, we can not only fully utilize the information in shape priors but also overcome the shortcomings of attention-based fusion methods to better reconstruct 3D objects.

In conclusion, the main contributions are summarized as follows:

- Tbject shapes highly related to the input image and extract useful knowledge from them to form shape prior vectors. By adopting cross-modal attention mechanisms, image and shape prior information can be effectively fused together and

Manuscript received March 11, 2024

**Xiaobing Zhang** , School of computer science and technology, Tiangong University, Tianjin, China

**Tengfei Xiao,** School of computer science and technology, Tiangong University, Tianjin, China.

forwarded to the decoder to predict the 3D shape of the object.

- Experimental results show that the proposed RSP3D method significantly improves the reconstruction quality on the ShapeNet and Pix3D datasets and outperforms the state of-the-art methods.

## II. RELATED WORKS

- Single-view 3D reconstruction. In recent years, estimating 3D shapes from single-view images has attracted a lot of attention. PointSetGeneration generates 3D shapes based on point cloud representations. Pixel2Mesh represents objects with triangle meshes and processes them with graph convolutional networks (GCN). Voxel representations are common. 3DCNN is used directly to modify voxel grids. Generative adversarial networks (GANs) propose 3DGAN and 3DIWGAN to solve 3D object generation problems, and combine variational autoencoders (VAEs) to convert these works' generators into single-view reconstructors. For high-resolution results, OGN adopts an octree representation to overcome huge memory budget issues and designs a network to process it directly. However, Matryoshka Networks recursively decompose 3D shapes into nested shape layers. To bridge the gap between synthetic and real-world data, DAREC and VPAN introduce domain-adaptive supervision during training. To supplement missing information in images, Mem3D builds a memory network to provide accumulated prior information from a training set.

- Multi-view 3D Reconstruction. Traditional reconstruction methods such as SFM and SLAM rely on matching features to establish relationships between different views, but they have significant limitations in practical applications. Recently, methods based on deep learning have become popular for multi-view 3D reconstruction, typically without the need for viewpoint labels. This approach utilizes 2D CNNs to predict dense point clouds representing 3D object surfaces. In Pixel2Mesh++, coarse meshes can be iteratively improved through a series of deformations predicted by GCN to produce final results. This method utilizes voxel representations and focuses on how to merge features from multiple views. Max pooling layers are used to compress concatenated features from all views. The 3D-R2N2 series and LSM receive views one by one through recursive units and extract useful knowledge. EVolT and LegoFormer leverage the advantages of transformer architectures to achieve information fusion between various views during the encoder stage. As the most stable method currently, AttSets and Pix2Vox series apply attention modules to multi-branch tasks, but lack information exchange between branches.

- Memory Network. The Memory Network was first proposed in , which enhances the neural network with external memory modules to enable the network to store long-term memories. Subsequent work has improved the memory network to enable end-to-end training. Hierarchical Memory Networks have been proposed, which allow the reading controller to access large-scale memory effectively. Key-Value Memory Networks store prior knowledge in a key-value structured memory, where keys are used to address and retrieve relevant memories with corresponding values

## III. METHOD

In existing single-view 3D reconstruction methods [36,28, 37, 4], the shape priors are learnt into model parame-ters, which leads to low quality reconstructions for images containing heavy occlusion and noisy backgrounds. To alleviate this issue, the proposed RSP3D explicitly constructs the shape priors using a Key-Value Memory Network [17]. Specifically, the image encoder extracts features from the input image. During training, the extracted features and the corresponding 3D shape are then stored in the memory network in a key-value fashion. For both training and testing, the 3D shapes whose corresponding keys have high similarities are forwarded to the LSTM shape encoder. After that, the LSTM shape encoder generates a shape prior vector. Finally, the decoder takes the both image features and the shape prior vector to reconstruct the 3D shape of the object.

### *A. Memory Network*

The memory network aims to explicitly construct the shape priors by storing the "image-voxel" pairs, which memorize the correspondence between the image features and the corresponding 3D shapes.The memory items are constructed as: [key, value, age], which is denoted as M = {(Ki,Vi, Ai)m i=1}, where m denotes the size of the memory. The "key" and "value" memory slots store the image features and the corresponding 3D volume, respectively. The "key" Ki ∈ $R^{nk}$ is used to compute the cosine similar ities with the input image features. The "value" Vi ∈ $R^{nv}$ is returned if the similarity score between the query and thekeys of memory exceeds a threshold. The nk and nv are dimension of the memory "key" and memory "value", respectively. The "age" Ai ∈ N represents the alive time of the pair, which is to set to zero when the pair is matched by the input image features. The memory network overwrites the "oldest" pair when writing new pairs.

### *B. LSTM Shape Encoder*

The value sequence V retrieved by the memory readercontains 3D shapes that are similar to the object in the input image. The value sequence from the memory reader is length-variant and has been ordered by the similarities. Intuitively, different parts of different shapes in the value sequence may have a different importance in reconstructing the 3D shape from the current image. To contextually consider and incorporate knowledge useful for current reconstruction from the value sequence into the image feature to supplement the occluded or noisy parts, we leverage LSTM [9] to encode the value sequence V in a sequential manner. The LSTM shape encoder takes the length-variant value sequence as input and outputs a fixed-length "shape prior vector". The "shape prior vector" is then concatenated with the input image feature to provide extra useful information for the shape decoder.

### *C. Network Architecture*

**Image Encoder.** The image encoder contains the first threeconvolutional blocks of ResNet-50 [8] to extract a $512\times28^2$ feature map from a 224 × 224 × 3 image. Then the ResNet is followed by three sets of 2D convolutional layers, batch normalization layers and ReLU layers. The kernel sizes of the three convolutional layers are 32, with a padding of 1.

There is a max pooling layer with a kernel size of 22 after the second and third ReLU layers. The output channels of the three convolutional layers are 512, 256, and 256, respectively.

**LSTM Shape Encoder.** The shape encoder is an LSTM [9]network with 1 hidden layer. The hidden size is set to 2,048 which indicates that the output shape prior vector is a 2,048 dimensional vector.

**Shape Decoder**. The decoder contains five 3D transposed convolutional layers. The first four transposed convolutional layers are of kernel sizes 43, with strides of 2 and paddings of 1. The next transposed convolutional layer has a bank of 13 filter. Each of the first four transposed convolutional layers is followed by a batch normalization layer and a ReLU, and the last transposed convolutional layer is followed by a sigmoid function. The output channel numbersof the five transposed convolutional layers are 512, 128, 32,8, and 1, respectively. The final output of decoder is a $32^3$voxelized shape.

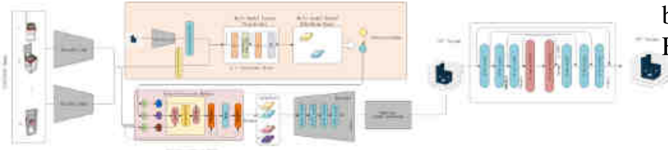


Figure 1. The proposed RSP3D reconstruct the 3D shape of an object from a single input image. The Memory Network learns to retrieve 3D volumes that are highly related to the input image. The LSTM Shape Encoder is proposed to contextually encode multiple 3D volumes into a shape prior vector, which provides the information that helps to recover the 3D shape of the object's hidden and noisy parts.

## IV. EXPERIMENTS

### A. Datasets

**ShapeNet.** The ShapeNet dataset [2] is composed of synthetic images and corresponding 3D volumes. We use a subset of the ShapeNet dataset consisting of 44K models and 13 major categories following [4]. Specifically, we use renderings provided by 3D-R2N2 which contains 24 random views of size 137 × 137 for each 3D model. We also apply random background augmentation [36, 22] to the image during training. Note that only the ShapeNet dataset is used for training Mem3D.

**Pix3D.** The Pix3D [24] dataset contains 395 3D models of nine classes. Each model is associated with a set of real images, capturing the exact object in diverse environments. The most significant category in this dataset is chairs. The Pix3D dataset is used only for evaluation.

### B. Evaluation Metrics

We apply the intersection over union (IoU) and F-score evaluation metrics widely used by existing works. The IoU is formulated as

$$\text{IoU} = \frac{\sum_{i,j,k} \mathbb{I}(p(i,j,k) > t)\mathbb{I}(gt(i,j,k))}{\sum_{i,j,k} \mathbb{I}[\mathbb{I}(p(i,j,k) > t) + \mathbb{I}(gt(i,j,k))]}$$

where p(i, j, k) and gt(i, j, k) indicate predicted occupancy probability and ground-truth at (i,j,k), respectively. I is thindication function which will equal to one when the requirements are satisfied. The t denotes a threshold, t = 0.3 in our experiments. Following Tatarchenko et al. [26], we also take F-Score as an extra metric to evaluate the performance of 3D reconstruction results, which can be defined as

$$\text{F-Score}(d) = \frac{2P(d)R(d)}{P(d) + R(d)}$$

where P(d) and R(d) denote the precision and recall with adistance threshold d, respectively. P(d) and R(d) are computed as

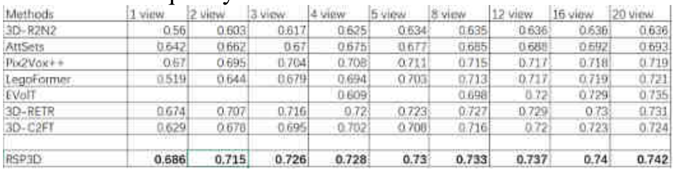
$$P(d) = \frac{1}{n_{\mathcal{R}}} \sum_{r \in \mathcal{R}} \left[ \min_{g \in \mathcal{G}} ||g - r|| < d \right]$$

$$R(d) = \frac{1}{n_{\mathcal{G}}} \sum_{g \in \mathcal{G}} \left[ \min_{r \in \mathcal{R}} ||g - r|| < d \right]$$

where R and G represent the predicted and ground truth point clouds, respectively. nR and nG are the number of points in R and G, respectively. To adapt the F-Score to voxel models, like existing works [36], we apply the marching cube algorithm [13] to generate the object surface, then 8,192 points are sampled from the surface to compute F-Score between predicted and ground truth voxels. A higher IoU and F-Score indicates better reconstruction results.

### C. Experiment Result

We compare the performance with other state-of-the- art methods on the ShapeNet testing set. Tables 1 and 2 show the IoU and F-Score@1% of all methods, respectively, which indicates that Mem3D outperforms all other competitive methods with a large margin in terms of both IoU and F-Score@1%. Our Mem3D benefits from the memory network which explicitly constructs shape priors and applies them according to an object's individual needs to improve reconstruction quality.



| Methods | 1 view | 2 view | 3 view | 4 view | 5 view | 8 view | 12 view | 16 view | 20 view |
|---|---|---|---|---|---|---|---|---|---|
| 3D-R2N2 | 0.56 | 0.603 | 0.617 | 0.625 | 0.634 | 0.635 | 0.636 | 0.636 | 0.636 |
| AttSets | 0.642 | 0.662 | 0.67 | 0.675 | 0.677 | 0.685 | 0.688 | 0.692 | 0.693 |
| Pix2Vox++ | 0.67 | 0.695 | 0.704 | 0.708 | 0.711 | 0.715 | 0.717 | 0.718 | 0.719 |
| LegoFormer | 0.519 | 0.644 | 0.679 | 0.694 | 0.703 | 0.713 | 0.717 | 0.719 | 0.721 |
| EVolT |  |  |  | 0.609 |  | 0.698 | 0.72 | 0.729 | 0.735 |
| 3D-RETR | 0.674 | 0.707 | 0.716 | 0.72 | 0.723 | 0.727 | 0.729 | 0.73 | 0.731 |
| 3D-C2FT | 0.629 | 0.678 | 0.695 | 0.702 | 0.708 | 0.716 | 0.72 | 0.723 | 0.724 |
|  |  |  |  |  |  |  |  |  |  |
| RSP3D | **0.686** | **0.715** | **0.726** | **0.728** | **0.73** | **0.733** | **0.737** | **0.74** | **0.742** |

Table 1. Evaluation and comparison of the performance on ShapeNet using IoU / F-Score@1%. The best results are highlighted in bold.

## V. CONCLUSION

In this paper, we propose a novel framework for 3D object reconstruction, named RSP3D. Compared to the existing methods for single-view and mutil-view 3D object reconstruction that directly learn to transform image features into 3D representations, RSP3D constructs shape priors that are helpful to complete the missing image features to recover the 3D shape of an object that is heavy occluded or in a complex environment. Experimental results demonstrate that RSP3D significantly improves the reconstruction quality and performs favorably against state-of-the-art methods on the ShapeNet and Pix3D datasets.

## REFERENCES

[1] Sarath Chandar, Sungjin Ahn, Hugo Larochelle, Pascal Vincent, Gerald Tesauro, and Yoshua Bengio. Hierarchical memory networks. CoRR, 2016. 2

[2] Angel X. Chang, Thomas A. Funkhouser, Leonidas J Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong

Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. CoRR, 2015. 1, 5

[3] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In CVPR, 2019. 6

[4] Christopher Bongsoo Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In ECCV, 2016. 1, 2, 5, 6, 7

[5] Endri Dibra, Himanshu Jain, Cengiz Oztireli, Remo Ziegler, and Markus Gross. Human shape from silhouettes using generative hks descriptors and cross-modal neural networks. In CVPR, 2017. 2

[6] Haoqiang Fan, Hao Su, and Leonidas J. Guibas. A point set generation network for 3d object reconstruction from a single image. In CVPR, 2017. 2

[7] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In NIPS, 2014. 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016. 4

[9] Sepp Hochreiter and J¨urgen Schmidhuber. Long short-term memory. Neural Computation, 9(8):1735–1780, 1997. 4, 8

[10] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In ICLR, 2014. 2

[11] Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer,James Bradbury,Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. In ICML, 2016. 2

[12] Abhijit Kundu, Yin Li, and James M. Rehg. 3d-rcnn:Instance-level 3d object reconstruction via render-and-compare. In CVPR, 2018. 2

[13] William E. Lorensen and Harvey E. Cline. Marching cubes:A high resolution 3d surface construction algorithm. In SIGGRAPH, 1987. 5

[14] Priyanka Mandikal, Navaneet K. L., Mayank Agarwal, and Venkatesh Babu Radhakrishnan. 3d-lmnet: Latent embedding matching for accurate and diverse 3d point cloud reconstruction from a single image. In BMVC, 2018. 2

[15] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks:Learning 3d reconstruction in function space. In CVPR,2019. 6

[16] Alexander H. Miller, Adam Fisch, Jesse Dodge, AmirHossein Karimi, Antoine Bordes, and Jason Weston. Key-value memory networks for directly reading documents. In EMNLP, 2016. 2

[17] Pedro O. Pinheiro, Negar Rostamzadeh, and Sungjin Ahn.Domain-adaptive single-view 3d reconstruction. In ICCV,2019. 7

[18] Stephan R. Richter and Stefan Roth. Discriminative shape from shading in uncalibrated illumination. In CVPR, 2015.2

[19] Martin R¨unz, Kejie Li, Meng Tang, Lingni Ma, Chen Kong, Tanner Schmidt, Ian D. Reid, Lourdes Agapito, Ju-lian Straub, Steven Lovegrove, and Richard A. Newcombe.Frodo: From detections to 3d objects. In CVPR, 2020. 2, 7

[20] Florian Schroff, Dmitry Kalenichenko, and James Philbin.Facenet: A unified embedding for face recognition and clustering. In (CVPR), 2015. 4