Single-View 3D Object Reconstruction from Shape Priors in Memory

Shuo Yang[1] Min Xu[1*] Haozhe Xie[2] Stuart Perry[1] Jiahao Xia[1]
[1] School of Electrical and Data Engineering, University of Technology Sydney
[2] Harbin Institute of Technology
{shuo.yang,jiahao.xia}@student.uts.edu.au
{min.xu,stuart.perry}@uts.edu.au cshzxie@gmail.com

Abstract

Existing methods for single-view 3D object reconstruction directly learn to transform image features into 3D representations. However, these methods are vulnerable to images containing noisy backgrounds and heavy occlusions because the extracted image features do not contain enough information to reconstruct high-quality 3D shapes. Humans routinely use incomplete or noisy visual cues from an image to retrieve similar 3D shapes from their memory and reconstruct the 3D shape of an object. Inspired by this, we propose a novel method, named Mem3D, that explicitly constructs shape priors to supplement the missing information in the image. Specifically, the shape priors are in the forms of "image-voxel" pairs in the memory network, which is stored by a well-designed writing strategy during training. We also propose a voxel triplet loss function that helps to retrieve the precise 3D shapes that are highly related to the input image from shape priors. The LSTM-based shape encoder is introduced to extract information from the retrieved 3D shapes, which are useful in recovering the 3D shape of an object that is heavily occluded or in complex environments. Experimental results demonstrate that Mem3D significantly improves reconstruction quality and performs favorably against state-of-the-art methods on the ShapeNet and Pix3D datasets.

1. Introduction

Reconstructing object 3D shape from a single-view RGB image is a vital but challenging computer vision task in robotics, CAD, and virtual and augmented reality applications. Humans can easily infer the 3D shape of an object from a single image due to sufficient prior knowledge and an innate ability for visual understanding and reasoning. However, this is an extremely difficult and ill-posed problem for a machine vision systems because a single-view image can not provide sufficient information for the object to be reconstructed.

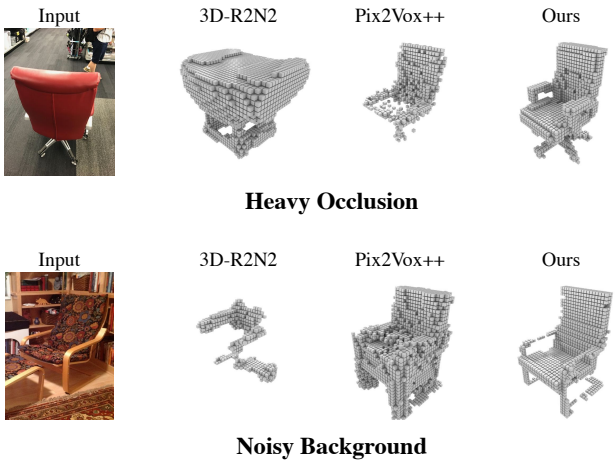

Figure 1. Compared with the current state-of-the-arts method Pix2Vox++ [36] and classic method 3D-R2N2 [4], the proposed method are more robust in reconstructing the 3D shape of an object from a single image that contains occlusion or noisy backgrounds.

Most of the existing learning-based methods for single-view 3D reconstruction extract features from a single RGB image, then transform it into a 3D representation. These methods achieve promising results on the synthetic datasets (ShapeNet [2]). However, as shown in Figure 1, they usually have trouble reconstructing the 3D shape of an object from real-world images. The performance gap between the real-world and synthetic datasets are caused by the quality of image features. The features extracted from images with noisy backgrounds or heavy occlusions usually contain insufficient useful information for 3D reconstruction.

Humans can infer a reasonable 3D shape of an object from a single image, even with incomplete or noisy visual cues. This is due to the fact that humans retrieve similar shapes from their memories and apply these shape priors to recover the shape of hidden and noisy parts of the object. Motivated by human vision, we propose a novel memory-

*Corresponding Author.

based framework for 3D reconstruction, Mem3D, which consists of four components: image encoder, memory network, LSTM shape encoder, and shape decoder. The proposed Mem3D explicitly constructs the shape priors in the memory network that help complete the missing image features to recover the 3D shape of an object that is heavy occluded or in a complex environment. To construct the shape priors, we design a writing strategy to store the "image-voxel" pairs into the memory network in a key-value fashion during training. To retrieve the precise 3D shapes that are highly related to the input image from the memory network, we propose a voxel triplet loss function that guarantees that images with similar 3D shapes are closer in the feature space. To better leverage the retrieved shapes, the LSTM-based shape encoder transforms the useful knowledge of these shapes into a shape prior vector. To employ both information from image and shape priors, the input image features and the output of the LSTM-based shape encoder are concatenated and are forwarded to a decoder to predict the 3D shape of the object.

The main contributions are summarized as follows:

- We propose a memory-based framework for single-view 3D object reconstruction, named Mem3D. It innovatively retrieves similar 3D shapes from the constructed shape priors, and shows a powerful ability to reconstruct the 3D shape of objects that are heavily occluded or in a complex environment.
- We present a memory network that stores shape priors in the form of "image-voxel" pairs. To better organize the shape priors and ensure accurate retrieval, we design novel reading and writing strategies, as well as introducing a voxel triplet loss function.
- Experimental results demonstrate that the proposed Mem3D significantly improves the reconstruction quality and performs favorably against state-of-the-art methods on the ShapeNet and Pix3D datasets.

2. Related Work

Single-image 3D Reconstruction. Recently, 3D object reconstruction from a single-view image has attracted increasing attention because of its wide applications in the real world. Recovering object shape from a single-view image is an ill-posed problem due to the limitation of visual clues. Existing works use the representation of silhouettes [5], shading [19], and texture [30] to recover 3D shape. With the success of deep learning, especially generative adversarial networks [7] and variational autoencoders [10], the deep neural network based encoder-decoder has become the main-stream architecture, such as 3D-VAE-GAN [32]. PSGN [6] and 3DLMNet [14] generate point representations from single-view images. 3D-R2N2 [4] is a unified framework for single- and multi-view 3D reconstruction which employs a 3D convolutional LSTM to fuse the image features. To solve the permutation variance issue, Pix2Vox [35] employs a context-aware fusion module to adaptively select high-quality reconstructions from single-view reconstructions. However, these works that utilize shape priors implicitly are venerable to noisy backgrounds and heavy occlusions. To reconstruct the 3D shape of an object from real-world images, MarrNet [31] and its variants [33, 40] reconstructs 3D objects by estimating depth, surface normals, and silhouettes. Both 3D-RCNN [12] and FroDO [20] introduce an object detector to remove noisy backgrounds.

Memory Network. The Memory Network was first proposed in [29], which augmented neural networks with an external memory module that enables the neural network to store long-term memory. Later works [11, 23] improve the Memory Network so it can be trained in an end-to-end manner. Hierarchical Memory Networks [1] was proposed to allow a read controller to efficiently access large scale memories. Key-Value Memory Networks [16] store prior knowledge in a key-value structured memory, where keys are used to address relevant memories whose corresponding values are returned.

3. Method

In existing single-view 3D reconstruction methods [36, 28, 37, 4], the shape priors are learnt into model parameters, which leads to low quality reconstructions for images containing heavy occlusion and noisy backgrounds. To alleviate this issue, the proposed Mem3D 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.

3.1. 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 $M = \{(\mathbf{K}_i, \mathbf{V}_i, A_i)_{i=1}^{m}\}$, 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" $\mathbf{K}_i \in \mathbb{R}^{n_k}$ is used to compute the cosine similar-

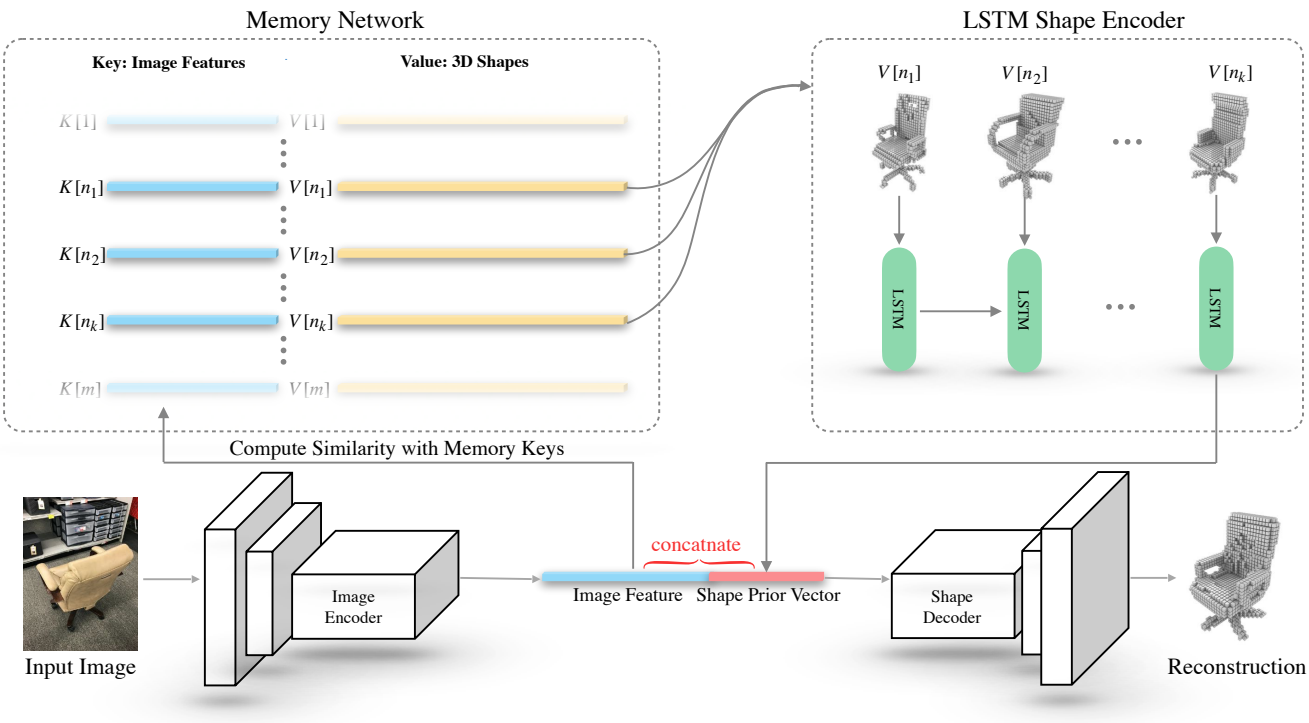

Figure 2. The proposed Mem3D reconstruct the 3D shape of an object from a single input image. The Memory Network learns to retrieve 3D volumes that are highly related to the input image. The LSTM Shape Encoder is proposed to contextually encode multiple 3D volumes into a shape prior vector, which provides the information that helps to recover the 3D shape of the object's hidden and noisy parts.

ities with the input image features. The "value" $\mathbf{V}_i \in \mathbb{R}^{n_v}$ is returned if the similarity score between the query and the keys of memory exceeds a threshold. The n_k and n_v are dimension of the memory "key" and memory "value", respectively. The "age" $\mathbf{A}_i \in \mathbb{N}$ 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.

3.1.1 Memory Writer

The memory writer is presented to construct the shape priors in the memory network. We designed a writing strategy to determine how to update the memory slots when given the image features $\mathbf{F} \in \mathbb{R}^{n_k}$ and its corresponding volume $\mathcal{V}$. The memory writing only works at training because it takes the ground truth 3D volumes as input, which are not available during testing. In the memory network, the key similarity between the input image features $\mathbf{F}$ and the memory key $\mathbf{K}_i$ is defined as following

$$S_k(\mathbf{F}, \mathbf{K}_i) = \frac{\mathbf{F} \cdot \mathbf{K}_i}{\|\mathbf{F}\|\|\mathbf{K}_i\|} \tag{1}$$

Similarly, the value similarity between the corresponding 3D volumes $\mathcal{V}$ and the value $\mathbf{V}_i$ can be defined as

$$S_v(\mathcal{V}, \mathbf{V}_i) = 1 - \frac{1}{r_v^3} \sum_{j=1}^{r_v^3} (\mathbf{V}_i^j - \mathcal{V}^j)^2 \tag{2}$$

where r_v indicates the resolution of the 3D volume.

The writing strategy works for the two cases according to whether the similarity satisfies $S_v(\mathcal{V}, \mathbf{V}_{n_1}) > \delta$, where δ is the similarity threshold and n_1 is determined by

$$n_1 = \arg\max_i S_k(\mathbf{F}, \mathbf{K}_i) \tag{3}$$

Strategy for Similar Examples ($S_v(\mathcal{V}, \mathbf{V}_{n_1}) \geq \delta$). For a similar example, the value $\mathbf{V}_{n_1}$ kept unchanged, while the age $A_{n_1} = 0$ and the key $\mathbf{K}_{n_1}$ is updated as follows:

$$\mathbf{K}_{n_1} = \frac{\mathbf{F} + \mathbf{K}_{n_1}}{\|\mathbf{F} + \mathbf{K}_{n_1}\|} \tag{4}$$

After the memory update, the ages are adjusted as $A_i = A_i + 1$ $(i \neq n_1)$.

Strategy for New Examples ($S_v(\mathcal{V}, \mathbf{V}_{n_1}) < \delta$). For a new example, the memory writer stores it to the memory network as following

$$\mathbf{K}_{n_o} = \mathbf{F} \tag{5}$$

$$\mathbf{V}_{n_o} = \mathcal{V} \tag{6}$$

$$A_{n_o} = 0 \tag{7}$$

where n_o is determined by

$$n_o = \arg\max_i(A_i) \tag{8}$$

if there are no empty slots in the memory network. Otherwise, n_o can be the index of any empty memory slots. After the memory update, the ages are adjusted as $A_i = A_i + 1$ $(i \neq n_o)$.

3.1.2 Memory Reader

The memory reader is used for reading the values from the memory network and outputs a value sequence containing 3D volumes that are highly related to the input image features. For different input image features, there are different numbers of highly similar shapes in the memory. Therefore, retrieving a fixed number of shapes from the memory network would not be suitable for all inputs and may introduce irrelevant shapes.

To solve this problem, we construct the retrieved value sequence by concatenating all values whose key satisfies $S_k(\mathbf{F}, \mathbf{K}_{n_i}) > \beta$, which can be formulated as

$$\mathbb{V} = [V_{n_i} | S_k(\mathbf{F}, \mathbf{K}_{n_i}) > \beta] \tag{9}$$

where β is threshold and $[\cdot]$ denotes the concatenation.

3.2. LSTM Shape Encoder

The value sequence $\mathbb{V}$ retrieved by the memory reader contains 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 $\mathbb{V}$ 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.

3.3. Network Architecture

Image Encoder. The image encoder contains the first three convolutional blocks of ResNet-50 [8] to extract a 512×28^2 feature map from a $224 \times 224 \times 3$ 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 3^2, with a padding of 1. There is a max pooling layer with a kernel size of 2^2 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 4^3, with strides of 2 and paddings of 1. The next transposed convolutional layer has a bank of 1^3 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 numbers of the five transposed convolutional layers are 512, 128, 32, 8, and 1, respectively. The final output of decoder is a 32^3 voxelized shape.

3.4. Loss Functions

Voxel Triplet Loss. We propose a voxel triplet loss that helps to retrieve precise values from the memory network by guaranteeing that images with similar 3D shapes are closer in the feature space. In the memory network, n_p and n_b are the memory slots of the positive and negative samples, respectively. For a positive sample, the similarity between its value $\mathbf{V}_{n_p}$ and the corresponding 3D volume of the input image $\mathcal{V}$ satisfies

$$S_v(\mathcal{V}, \mathbf{V}_{n_p}) \geq \delta \tag{10}$$

Similarly, for a negative sample, the similarity satisfies

$$S_v(\mathcal{V}, \mathbf{V}_{n_b}) < \delta \tag{11}$$

where $\mathbf{V}_{n_b}$ represents the value of the "image-voxel" pair for n_b. Therefore, the voxel triplet loss can be defined as

$$\ell_t(S_{kb}, S_{kp}, \alpha) = \max\left(S_{kb} - S_{kp} + \alpha, 0\right) \tag{12}$$

where α is the margin in the triplet loss [21, 39, 38]. S_{kb} and S_{kp} are the similarities between the input image features and the keys of the postive/negative sample, which are defined as $S_{kb} = S_k(\mathbf{F}, \mathbf{K}_{n_b})$ and $S_{kp} = S_k(\mathbf{F}, \mathbf{K}_{n_p})$, respectively. The proposed voxel triplet loss can minimize the distance among image features with similar 3D volumes and maximize the distance among image features with different 3D volumes.

Binary Cross Entropy Loss. For the reconstruction network, we adopt the Binary Cross Entropy Loss, which is defined as the mean value of the voxel-wise binary cross entropies between the reconstructed object and the ground truth. More formally, it can be defined as

$$\ell_r(p, gt) = \frac{1}{r_v^3} \sum_{i=1}^{r_v^3} [gt_i \log(p_i) + (1 - gt_i) \log(1 - p_i)] \tag{13}$$

where p and gt denote the predicted 3D volume and the corresponding ground truth, respectively.

The Mem3D is trained end-to-end by the combination of the voxel triplet loss and the reconstruction loss:

$$\ell_{total} = \ell_t + \ell_r \tag{14}$$

4. Experiments

4.1. 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.

4.2. Evaluation Metrics

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

$$\text{IoU} = \frac{\sum_{i,j,k} \mathbb{I}(p(i,j,k) > t)\mathbb{I}(gt(i,j,k))}{\sum_{i,j,k} \mathbb{I}[\mathbb{I}(p(i,j,k) > t) + \mathbb{I}(gt(i,j,k))]} \tag{15}$$

where $p(i,j,k)$ and $gt(i,j,k)$ indicate predicted occupancy probability and ground-truth at *(i,j,k)*, respectively. $\mathbb{I}$ is the indication 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

$$\text{F-Score}(d) = \frac{2P(d)R(d)}{P(d) + R(d)} \tag{16}$$

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

$$P(d) = \frac{1}{n_{\mathcal{R}}} \sum_{r \in \mathcal{R}} \left[\min_{g \in \mathcal{G}} ||g - r|| < d \right] \tag{17}$$

$$R(d) = \frac{1}{n_{\mathcal{G}}} \sum_{g \in \mathcal{G}} \left[\min_{r \in \mathcal{R}} ||g - r|| < d \right] \tag{18}$$

where $\mathcal{R}$ and $\mathcal{G}$ represent the predicted and ground truth point clouds, respectively. $n_{\mathcal{R}}$ and $n_{\mathcal{G}}$ are the number of points in $\mathcal{R}$ and $\mathcal{G}$, 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.

4.3. Implementation Details

We used 224×224 RGB images as input to train the Mem3D with a batch size of 32. The whole network is trained end-to-end with the Adam optimizer with a β_1 of 0.9 and a β_2 of 0.999. The initial learning rate is set to 0.001 and decayed by 2 after 150 epochs. In the memory network, the size is $m = 4000$. The margin α in Equation (12) is set to 0.1. The thresholds β and δ in Equations (9) and (10) are set to 0.85 and 0.90, respectively. The source code will be publicly available.

4.4. Object Reconstruction on ShapeNet

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.

4.5. Object Reconstruction on Pix3D

Pix3D is a more challenging benchmark which contains diverse real-world images and corresponding shapes. In Pix3D, the 'chair' category contains 3,839 images and the corresponding 3D models, which are the largest category of the dataset. Due to the complicated environment in images, the objects are frequently occluded by surroundings or themselves. Therefore, most of the previous works [35, 24, 34] evaluate their approaches using the hand-selected 2,894 untruncated and unoccluded 'chair' images to guarantee their models can capture enough information from the images. However, although this avoids the occlusion problem to some extent by selecting unoccluded testing samples, the previous reconstruction models still perform imperfectly because of the complicated background.

Table 1. Comparison of single-view 3D object reconstruction on ShapeNet. We report the per-category and overall IoU at 32^3 resolution. The best results are highlighted in bold.

Category	3D-R2N2 [4]	OGN [25]	DRC [27]	Pixel2Mesh [28]	IM-Net [3]	AttSets [37]	Pix2Vox [35]	Mem3D
Airplane	0.513	0.587	0.571	0.508	0.702	0.594	0.674	**0.767**
Bench	0.421	0.481	0.453	0.379	0.564	0.552	0.608	**0.651**
Cabinet	0.716	0.729	0.635	0.732	0.680	0.783	0.799	**0.840**
Car	0.798	0.828	0.755	0.670	0.756	0.844	0.858	**0.877**
Chair	0.466	0.483	0.469	0.484	0.644	0.559	0.581	**0.712**
Display	0.468	0.502	0.419	0.582	0.585	0.565	0.548	**0.631**
Lamp	0.381	0.398	0.415	0.399	0.433	0.445	0.457	**0.535**
Speaker	0.662	0.637	0.609	0.672	0.683	0.721	0.721	**0.778**
Rifle	0.544	0.593	0.608	0.468	0.723	0.601	0.617	**0.746**
Sofa	0.628	0.646	0.606	0.622	0.694	0.703	0.725	**0.753**
Table	0.513	0.536	0.424	0.536	0.621	0.590	0.620	**0.685**
Cellphone	0.661	0.702	0.413	0.762	0.762	0.743	0.809	**0.823**
Watercraft	0.513	0.632	0.556	0.471	0.607	0.601	0.603	**0.684**
overall	0.560	0.596	0.545	0.552	0.659	0.642	0.670	**0.729**

Table 2. Comparison of single-view 3D object reconstruction on ShapeNet. We report the per-category and overall F-Score@1%. For voxel reconstruction methods, the points are sampled from triangular meshes generated by the marching cube algorithm. The best results are highlighted in bold.

Category	3D-R2N2 [4]	OGN [25]	OccNet [15]	Pixel2Mesh [28]	IM-Net [3]	AttSets [37]	Pix2Vox++ [36]	Mem3D
Airplane	0.412	0.487	0.494	0.376	0.598	0.489	0.583	**0.671**
Bench	0.345	0.364	0.318	0.313	0.361	0.406	0.478	**0.525**
Cabinet	0.327	0.316	0.449	0.450	0.345	0.367	0.408	**0.517**
Car	0.481	0.514	0.315	0.486	0.304	0.497	0.564	**0.590**
Chair	0.238	0.226	0.365	0.386	0.442	0.334	0.309	**0.503**
Display	0.227	0.215	0.468	0.319	0.466	0.310	0.296	**0.498**
Lamp	0.267	0.249	0.361	0.219	0.371	0.315	0.315	**0.403**
Speaker	0.231	0.225	0.249	0.190	0.200	0.211	0.152	**0.262**
Rifle	0.521	0.541	0.219	0.340	0.407	0.524	0.574	**0.626**
Sofa	0.274	0.290	0.324	0.343	0.354	0.334	0.377	**0.434**
Table	0.340	0.352	0.549	0.502	0.461	0.419	0.406	**0.569**
Cellphone	0.504	0.528	0.273	0.485	0.423	0.469	0.633	**0.674**
Watercraft	0.305	0.328	0.347	0.266	0.369	0.315	0.390	**0.461**
Overall	0.351	0.368	0.393	0.398	0.405	0.395	0.436	**0.517**

To show the superior ability to reconstruct objects with the occlusion and background issues, we evaluate Mem3D on the 2,894 chair images with less occlusions but complicated backgrounds and 945 chair images (the complementary set) with heavy occlusions.

Note that the Mem3D is trained on the ShapeNet "chair" training set and evaluate on the Pix3D chair set. Since the memory network only writes "image-voxel" pairs during training, the memory network only contains shape priors extracted from the ShapeNet dataset.

4.5.1 Reconstruction with Complicated Backgrounds

Table 3 shows the evaluation performance of Mem3D and other works on the 2,894 untruncated and unoccluded 'chair' images. Note that these methods use different types of extra information. For instance, MarrNet [31], DRC [27] and ShapeHD [34] use extra depth, surface normals and silhouettes information. The proposed Mem3D outperforms the state-of-the-art methods by a large margin in terms of both IoU and F-Score@1%. The reconstruction results of our Mem3D and previous state-of-the-art works 'Pix2Vox++' [36] and 'Pix3D' [24] are shown in Figure 3. Compared to 'Pix2Vox++' [36] which employs an encoder-decoder structure, Mem3D can produce more clean and complete reconstruction results. The reconstructions from Mem3D also provide more details compared to other models. The memory network in Mem3D can explicitly store and utilize 3D volumes thus providing the reconstruction network with more detailed information about the object and eliminating the background noise.

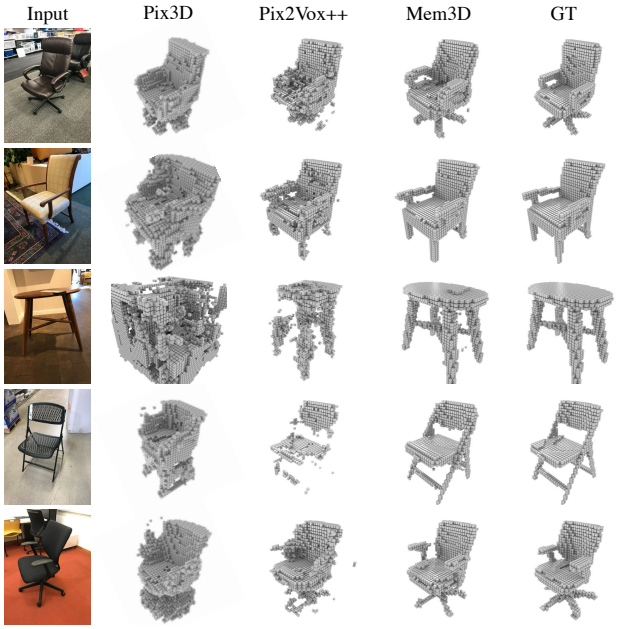

Figure 3. Reconstruction result on 5 of the 2,894 untruncated and unoccluded 'chairs' in Pix3D. GT indicates the ground-truth.

Table 3. Comparison of single-view 3D object reconstruction on 2,894 untruncated and unoccluded 'chair' images in Pix3D. We report IoU and F-Score@1%. The best performance is highlighted in bold.

Method	IoU	F-Score@1%
3D-R2N2[4]	0.136	0.018
3D-VAE-GAN [32]	0.171	-
MarrNet [31]	0.231	0.026
DRC [27]	0.265	0.038
ShapeHD [34]	0.284	0.046
DAREC [18]	0.241	-
Pix3D [24]	0.282	0.041
Pix2Vox++ [36]	0.292	0.068
FroDo [20]	0.325	-
Mem3D	**0.387**	**0.143**

4.5.2 Reconstruction with Heavy Occlusions

The occlusion issue is another key difficulty for single-view object reconstruction. Table 4 shows the evaluation performance of Mem3D and other works on the 945 chair images with heavy occlusions. The performance of all other methods drop significantly compared to Section 4.5.1. While Mem3D shows a favorable ability to handle the extremely cases. Figure 4 shows some reconstruction results of our Mem3D, 'Pix2Vox++' [36], and 3D-R2N2 [4]. It can be observed that Pix2Vox++ can reconstruct the perfectly presented parts of object in the image, but failed to reconstruct the occluded parts. Our Mem3D can provide reasonable reconstruction even for object parts that are hidden in the image. This is because Mem3D not only captures object information from images, but also obtains complete and clean shape information from the shapes read from memory. The retrieved shapes can provide detailed and complete shape information for the reconstruction network.

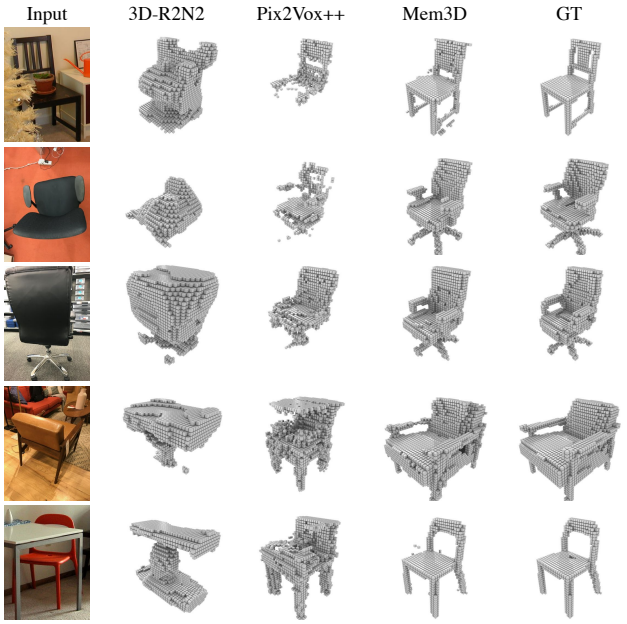

Figure 4. Reconstruction result on 5 of the 945 'chair' images with heavy occlusions in Pix3D. GT indicates the ground-truth.

Table 4. Comparison of single-view 3D object reconstruction on 945 'chairs' with heavy occlusions in Pix3D. We report IoU and F-Score@1%. The best performance is highlighted in bold.

Method	IoU	F-Score@1%
3D-R2N2[4]	0.055	0.011
MarrNet [31]	0.138	0.019
DRC [27]	0.151	0.025
ShapeHD [34]	0.183	0.037
Pix2Vox++ [36]	0.215	0.041
Mem3D	**0.336**	**0.105**

4.6. Ablation Study

In this section, we evaluate the importance of individual components by ablation studies.

Memory Network. To quantitatively evaluate the memory network, we remove the memory network and directly employ the image encoder and decoder as the baseline model. To further prove the effectiveness of the proposed voxel triplet loss ℓ_t in (12), which pulls image features with similar 3D shapes closer, we remove the voxel triplet loss ℓ_t from Mem3D training stage. The comparison results are shown in Table 5, which demonstrates the memory network and the proposed voxel triplet loss contribute significant improvement. We also show the reconstruction results of the baseline model (without the memory network) and our

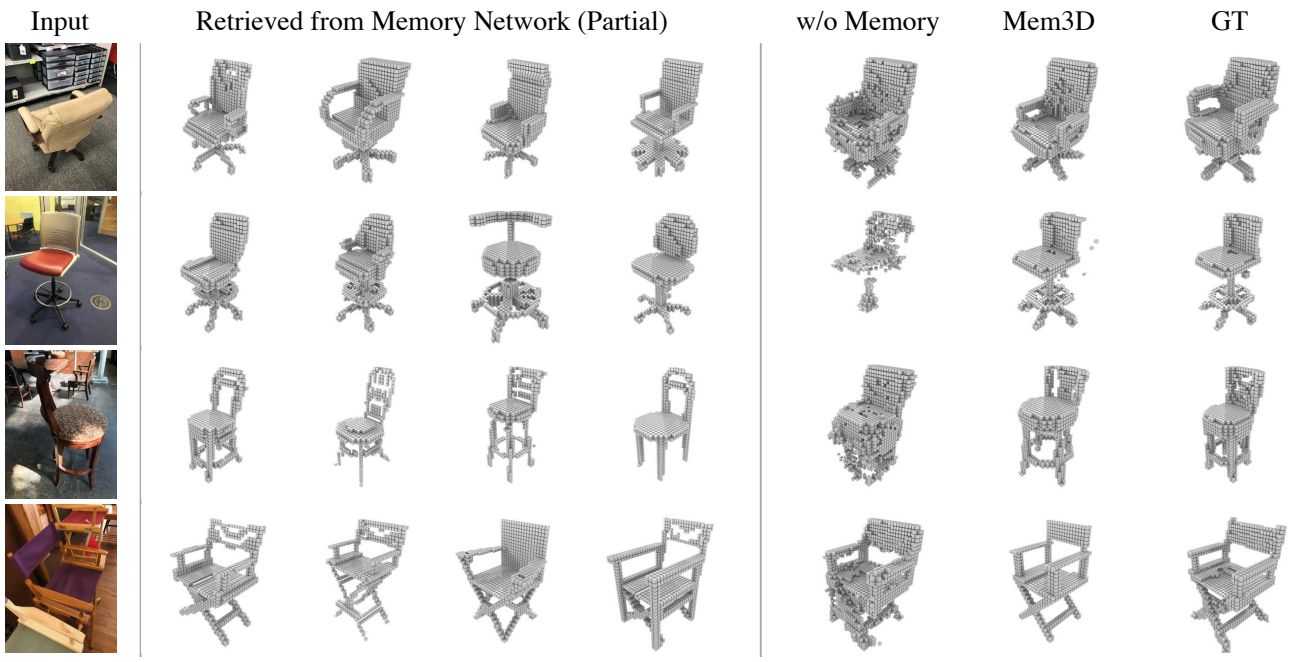

Figure 5. An illustration of the retrieved 3D volumes and the corresponding reconstructions. "w/o Memory'' indicates the reconstruction results are generated without the memory network. We only show the top-4 high-relative 3D volumes retrieved from the memory network. GT indicates the ground-truth.

Table 5. The effect of the memory network and the voxel triplet loss. The best results are highlighted in bold. 'm' indicates the memory size and ℓ_t indicates the voxel triplet loss in Equation (12).

Method	IoU	F-Score@1%
w/o memory network	0.273	0.042
m = 1000	0.366	0.113
m = 2000	0.372	0.135
m = 4000 w/o ℓ_t	0.359	0.111
m = 4000	**0.387**	**0.143**

Table 6. Different ways of leveraging retrieved shapes. 'Top-1' indicates directly treating the first retrieved shape as the reconstruction result. The best results are highlighted in bold.

Method	IoU	F-Score@1%
Top-1	0.287	0.051
Average Fusion	0.363	0.125
LSTM Shape Encoder	**0.387**	**0.143**

Mem3D as well as the retrieved shapes in Figure 5. The baseline model which reconstruct the 3D shape of an object from a single image captured in complicated environments are vulnerable to noisy backgrounds and occlusions. Our proposed Mem3D not only obtains object shapes from images, but also can access complete and clean relevant shapes during reconstruction. The proposed Mem3D makes it possible to reconstruct the hidden parts of the object in the image and significantly improve the reconstruction quality.

LSTM Shape Encoder. With the memory network in hand, we have different choices to leverage the retrieved shapes. For instance, we can directly use the Top-1 retrieved shape as reconstruction result, which is similar to retrieval-based reconstruction [26]. We can also use average fusion or the LSTM [9] network to encode useful knowledge from retrieved shapes into a fixed-length vector to condition the decoder. Table 6 shows the reconstruction performance when using different ways to leverage the retrieved shapes. We can also observe the retrieved shapes in Figure 5. The Top-1 retrieved shape has similar overall appearance compared to the ground-truth object shape, but the details are very different. The proposed Mem3D uses the image and the retrieved shapes together to provide high-quality reconstructions, which contains unique details that are distinct from the retrieved 3D volumes.

5. Conclusion

In this paper, we propose a novel framework for 3D object reconstruction, named Mem3D. Compared to the existing methods for single-view 3D object reconstruction that directly learn to transform image features into 3D representations, Mem3D 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 Mem3D 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 Jürgen 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, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. Key-value memory networks for directly reading documents. In *EMNLP*, 2016. 2

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

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

[19] Stephan R. Richter and Stefan Roth. Discriminative shape from shading in uncalibrated illumination. In *CVPR*, 2015. 2

[20] Martin Rünz, Kejie Li, Meng Tang, Lingni Ma, Chen Kong, Tanner Schmidt, Ian D. Reid, Lourdes Agapito, Julian Straub, Steven Lovegrove, and Richard A. Newcombe. Frodo: From detections to 3d objects. In *CVPR*, 2020. 2, 7

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

[22] Hao Su, Charles R. Qi, Yangyan Li, and Leonidas J. Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *(ICCV)*, 2015. 5

[23] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In *NIPS*, 2015. 2

[24] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B. Tenenbaum, and William T. Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *CVPR*, 2018. 5, 6, 7

[25] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *ICCV*, 2017. 6

[26] Maxim Tatarchenko, Stephan R. Richter, Rene Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *CVPR*, 2019. 5, 8

[27] Shubham Tulsiani, Tinghui Zhou, Alexei A. Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *CVPR*, 2017. 6, 7

[28] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*, 2018. 2, 6

[29] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. In *ICLR*, 2015. 2

[30] Andrew P. Witkin. Recovering surface shape and orientation from texture. *Artif. Intell.*, 17(1-3):17–45, 1981. 2

[31] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum. Marrnet: 3d shape reconstruction via 2.5d sketches. In *NIPS*, 2017. 2, 6, 7

[32] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *NIPS*, 2016. 2, 7

[33] Jiajun Wu, Chengkai Zhang, Xiuming Zhang, Zhoutong Zhang, William T Freeman, and Joshua B Tenenbaum. Learning 3D Shape Priors for Shape Completion and Reconstruction. In *ECCV*, 2018. 2

[34] Jiajun Wu, Chengkai Zhang, Xiuming Zhang, Zhoutong Zhang, William T. Freeman, and Joshua B. Tenenbaum.

Learning shape priors for single-view 3d completion and reconstruction. In *ECCV*, 2018. 5, 6, 7

[35] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *ICCV*, 2019. 2, 5, 6

[36] Haozhe Xie, Hongxun Yao, Shengping Zhang, Shangchen Zhou, and Wenxiu Sun. Pix2vox++: Multi-scale context-aware 3d object reconstruction from single and multiple images. *Int. J. Comput. Vis.*, 128(12):2919–2935, 2020. 1, 2, 5, 6, 7

[37] Bo Yang, Sen Wang, Andrew Markham, and Niki Trigoni. Robust attentional aggregation of deep feature sets for multi-view 3d reconstruction. *Int. J. Comput. Vis.*, 128(1):53–73, 2020. 2, 6

[38] Shuo Yang, Lu Liu, and Min Xu. Free lunch for few-shot learning: Distribution calibration. In *ICLR*, 2021. 4

[39] Shuo Yang, Wei Yu, Ying Zheng, Hongxun Yao, and Tao Mei. Adaptive semantic-visual tree for hierarchical embeddings. In *ACM MM*, 2019. 4

[40] Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Josh Tenenbaum, Bill Freeman, and Jiajun Wu. Learning to reconstruct shapes from unseen classes. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *NeurIPS 2018*, 2018. 2