CPG3D: Cross-Modal Priors Guided 3D Object Reconstruction

Weizhi Nie, Chuanqi Jiao, Rihao Chang, Lei Qu, and An-An Liu, *Senior Member, IEEE*

***Abstract*—Three-dimensional reconstruction is a multimedia technology widely used in computer-aided modeling and 3D animation. Nevertheless, it is still hard for reconstruction methods to overcome the 3D geometry missing and the object occlusion in the single-view images. In this article, we propose a novel method (CPG3D) for reconstructing high-quality 3D shapes from a single image under the guidance of prior knowledge. Using the single-view image as the query, prior knowledge is collected from public 3D datasets, which can compensate for missing 3D geometries and assist the 3D reconstruction network to high fidelity results. Our method consists of three parts: 1) Cross-modal 3D shape retrieval module: This part retrieves related 3D shapes based on 2D images. Here, we apply the pre-trained model to guarantee the correlation between the retrieved 3D shape and the input image. 2) Multimodal information fusion module: We propose a multimodal attention mechanism to handle the information fusing of 2D visual and 3D structural information; 3) Three-dimensional reconstruction module: We propose a novel encoder-decoder network for 3D shape reconstruction. Specifically, we employ the skip connection operation to link the target image's visual information with the 3D model's structural information to enhance the prediction of 3D details. During training, we employ two carefully designed loss functions to lead the multimodal learning to obtain proper modal features. On the ShapeNet and Pix3D datasets, the final experimental results reveal that our method notably increases reconstruction quality and outperforms SOTA methods.**

***Index Terms*—3D model reconstruction, Multimodal learning, Cross-modal retrieval.**

I. Introduction

THREE-DIMENSIONAL object reconstruction from a single-view RGB image is a widely researched multimedia task. It has been already applied in computer-assisted modeling [1], [2], 3D animation [3], [4] and robot localization [5], [6]. Unlike humans, who can easily infer the 3D shape of an object from a single image due to previously learned knowledge and an innate ability for visual understanding, computer vision systems are unable to directly reconstruct the 3D shapes from a single image because of the lack of structural information. Hence, the assistance of shape priors and the ability of shape reasoning play an indispensable and important role in 3D object reconstruction.

Lots of single-view 3D object reconstruction methods [7], [8], [9] rely on the massive training on large-scale datasets, which maps 2D image features into the 3D shape feature space through convolutional neural networks. However, this learning manner constrains their reconstruction ability due to the information loss in the single-view image. Although these methods have achieved some promising results on the views of 3D objects with clean backgrounds, they still fail to produce credible reconstruction results on views with complex backdrops, such as occlusion and truncation. This condition presents a new challenge to single-view 3D object reconstruction.

A. Motivation

Prior knowledge [10], [11], [12], [13] is an effective form to compensate for the 3D geometry loss, which provides the network with multimodal input for a comprehensive understanding of the target object. Consequently, the pipeline of prior-guided 3D reconstruction can be redesigned as: given a single-view image, computer vision systems should be able to automatically associate the shape prior knowledge and utilize this prior knowledge to guide the shape reconstruction. Thus, in this article, single-view reconstruction is divided into two tasks. 1) Obtaining the prior knowledge from the target image: Humans proactively learn the knowledge and memorize it as priors, while computers similarly obtain their priors from memory units. Therefore, this task constructs prior memories and retrieves the desired knowledge. 2) Reconstructing high-quality 3D shapes under the guidance of prior knowledge: Humans can infer the 3D structure of objects depicted in images based solely on prior knowledge. In a similar approach, computers should combine previously acquired priors with image data and infer the desired shape in a learnable manner. Consequently, this article focuses on efficient methods of information integration and shape reconstruction from the integrated features.

Based on previous analyses, we propose a novel 3D object reconstruction method that recovers the desired shape from a

Manuscript received 6 October 2022; revised 19 January 2023; accepted 20 February 2023. Date of publication 2 March 2023; date of current version 15 December 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFB1711704, in part by the National Archives Administration of China Science and Technology Project 2022-X-040 Design of 3D Electronic Document Intelligent Retrieval System Based on Data-based Thinking and Research on Key Technology Application, and in part by the Science and Technology Project of China Huaneng Group Company, Ltd. under Grant HNKJ22-H156 Design of 3D Electronic Document Intelligent Retrieval System Based on Data-based Thinking and Research on Key Technology Application. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Sebastian Knorr. *(Corresponding author: Rihao Chang.)*

Weizhi Nie, Chuanqi Jiao, and An-An Liu are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: weizhinie@tju.edu.cn; chuanqi_097@tju.edu.cn; anan0422@gmail.com).

Rihao Chang is with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China, and also with the School of Information Science and Engineering, Ocean University of China, Qingdao 266100, Shandong, China (e-mail: changrihao@tju.edu.cn).

Lei Qu is with the Hisense Group Holdings Company, Ltd., Qingdao 266000, China (e-mail: qulei1@hisense.com).

Digital Object Identifier 10.1109/TMM.2023.3251697

combined multimodal representation. First, to obtain appropriate shape priors, we utilize a cross-domain retrieval method that bridges the gap between images and 3D shapes. Given the input image, this method returns 3D shapes describing structurally homologous objects. For instance, when the network receives an image of a boss chair, the retrieval method will provide 3D chairs with structures such as large backrests and wheeled five-star feet. Second, we need to utilize the retrieved 3D shapes to make up for the missing structural information of the 2D target image. We first extract the features of images and 3D shapes under the supervision of carefully designed domain-specific loss functions, which learn the image features on the individual structures and the shape features on the overall structures, respectively. Then, we adopted the cross attention mechanism to explore the potential relation between the image feature and shape prior feature, based on which they are further combined into a highly integrated feature for 3D reconstruction. Finally, we expected the integrated feature to carry the overall homologous characteristics from priors and the differentiated individual characteristics from the single-view image. Thus, we adopted the 3D shape decoder to recover the 3D structural details from the integrated feature. Note that, to meet the previously desired reconstruction effect, the whole training process is under the supervision of the loss function that examines the reconstruction performance in 3D space.

B. Contributions

The contributions of this article are as follows:

- We propose a novel approach for reconstructing high-quality 3D shapes based on a given single-view image and its retrieved 3D prior knowledge. Shape priors can effectively compensate for the unseen structures in single-view images and optimize the performance of reconstruction;
- We propose a novel multimodal knowledge fusion network based on the cross attention mechanism, which can hierarchically fuse the image's visual information with relevant 3D shape structural information;
- We propose a three-dimensional reconstruction network assisted by the skip connections from image features. Besides, we validate the performance of the proposed method on the ShapeNet and Pix3D datasets. Several ablation studies are conducted to evaluate the effectiveness of each module. All experiments demonstrated the superiority and reasonableness of our network design.

The remainder of this article is organized as follows. Section II presents related works on 3D shape reconstruction and cross-modal 3D shape retrieval. Section III provides the details of our approach. The corresponding experimental results and analysis are described in Section IV. Finally, we discuss the contributions and conclude this article in Section V.

II. Related Work

A. Three-Dimensional Object Reconstruction

Due to the wide application prospects of 3D vision, the reconstruction of 3D objects from a single-view image has attracted increasing attention. The main concern of single-view reconstruction is how to recover the missing geometric information of the input image. Traditional methods dig deep into the potential 3D representation of the input image, such as shading [14], [15], [16], [17], occlusion [18], texture [19], [20] and vanishing points [21].

With the rapid development of deep learning and large-scale datasets, deep neural networks have shown their superiority. The early exploration, 3D-R2N2 [7], used a convolutional neural network with gated recurrent units to recover the full 3D structure of an object from single or several images. Unlike previous 3D reconstruction methods that rely on feature matching, 3D-R2N2 directly learns the mapping relationship from 2D to 3D space. However, 3D-R2N2 remains computationally inefficient. Inspired by the octree structure [22], [23], [24], researchers of HSP [25], OGN [26] and O-CNN [27] optimized the voxel representation by the voxel-level occupancy frequency and facilitated higher resolution prediction. However, it is still difficult for these methods to generate high-quality 3D shapes. Therefore, some researchers utilized the efficient 3D representations of the polygon mesh [28], [29], [30], [31], [32], [33] and deep implicit representation [34], [35], [36], [37], [38]. Polygon mesh represents the required shape with vertexes and faces, while implicit representation depicts the shape with a function between the 3D point position and the explicit or implicit distance from the point to the object surface. These methods can generate shapes with less redundant information. Besides, although the multi-view reconstruction task has long been dedicated to the combination of and consistency between different views, some methods [9], [39], [40] still have decent performance on single-view reconstruction tasks.

The key point of single-view reconstruction is the prior knowledge that compensates for the missing geometric information. Pontes et al. [31] represented the desired shape as a weighted combination of parameterized shape features. Yang et al. [11] first utilized memory networks [41], [42], [43], [44] to supplement 3D priors to 2D images and generate volumetric shapes. They introduced the LSTM-based shape encoder to extract shape priors from the memory network for reconstruction. Siddique et al. [12] leveraged the symmetries of 3D shapes as the prior knowledge. They considered that if the neighbors of an object partial are all symmetric, this partial is strongly judged as symmetric, and thus, the predicted shape can be optimized. Gao [13] et al. employed the domain adaptive learning for the alignment between 3D priors and 2D input images. They explored the multimodal correspondence between priors and input images, and further, they designed to divide the predicted shape according to the voxel grids and locally optimize the predicted shape. These previous works have verified the effectiveness of shape priors in 3D reconstruction. Thus, we employ cross-domain retrieval to precisely obtain desired 3D priors and design an efficient network to learn this guidance.

B. Image-Based 3D Object Retrieval

For input images, the compensation of the 3D structure information involves the challenge of cross-domain learning. Inspired

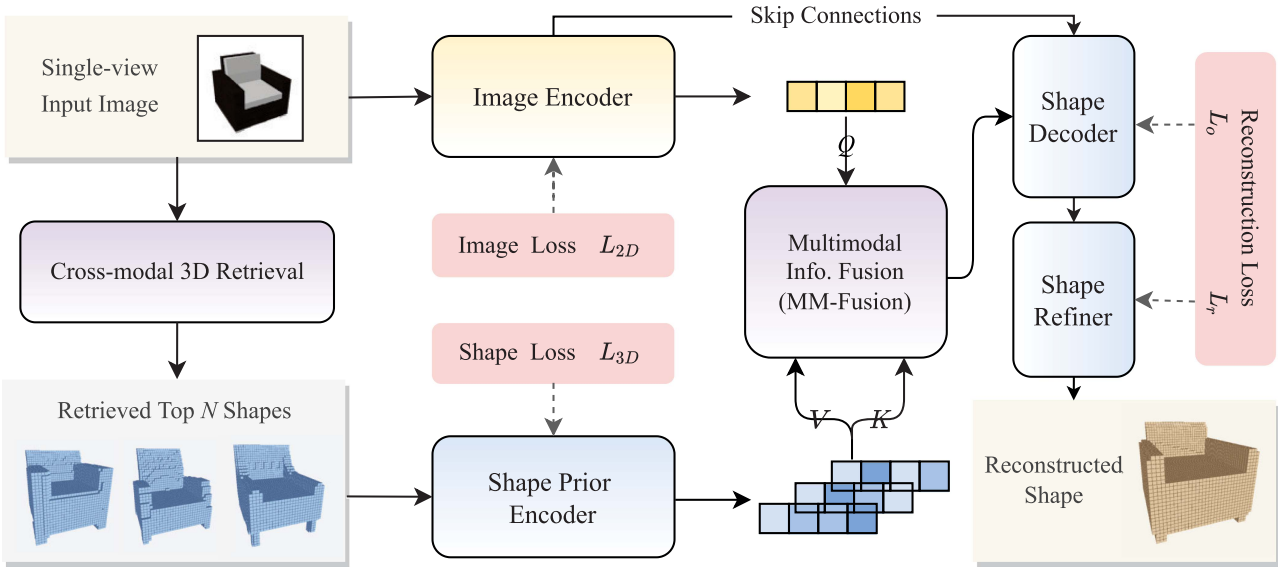


Fig. 1. The framework includes three parts: 1) Cross-modal prior knowledge extraction: input single-view image is used to retrieve the related 3D shape priors. 2) Multimodal information learning: the loss function L_{2D} and L_{3D} is utilized to supervise the feature extraction of the input image and shape priors, and these extracted features are further fused by the cross-modal attention modules. 3) Three-dimensional model reconstruction: The decoder is used to recover the 3D shape, and the refiner is adopted to optimize this reconstructed shape in detail.

by the prior-guided method of 3D reconstruction, cross-domain retrieval appears to be a more effective alternative to the acquisition of priors. Early methods [45], [46], [47] relied on a linear classifier to match the related shapes. However, the performance of these simple structures was unsatisfactory. Therefore, metric learning was introduced in many works to reinforce the performance of cross-domain retrieval. Lee et al. [48] proposed cross-view convolution to sequentially learn multi-view features for 3D shape representations. Moreover, they proposed a cross-domain triplet neural network to incorporate metric learning with cross-domain retrieval tasks. Mu et al. [49] described each input point as a point in Euclidean space and then mapped each 3D shape to a Riemannian manifold. To reduce the gap between the image and the 3D shape in different spaces, they bridged these two spaces to a shared high-dimensional Hilbert space, which greatly facilitated feature matching. Hence, cross-domain metric learning can effectively address the problem of complementing the 3D structure priors for the input image. More details about our network will be introduced in the following sections.

III. Approach

In this section, we will detail our approach. Fig. 1 shows the framework of our approach, which is divided into three parts.

1) Cross-modal prior knowledge extraction: Given the target image, a cross-modal retrieval approach is adopted to extract the shape prior knowledge. To guarantee the correlation of retrieved results with the target image, we apply a pre-trained CLN model [50], which will be introduced in the following subsection. Finally, we select the top N retrieved results as the 3D prior knowledge to compensate for the missing structural information.

2) Multimodal information learning: On the one hand, for each modality, the network extracts the individual 3D information from the input image and the common 3D information from the shape priors. On the other hand, the integration of individual information and common information is generated from our cross-modal attention modules.

3) Three-dimensional shape reconstruction: The 3D shape reconstruction is solved by adopting the decoder to recover the 3D information from the fused feature. Moreover, a shape refiner [40] is employed for better performance, but in the sacrifice of network lightness.

We will detail these three parts in the following subsections.

A. Cross-Modal Prior Knowledge Extraction

Following the previous work of [50], we deploy the cross-modal retrieval method based on metric learning. In [50], Nie et al. first adopted a pose estimation network to predict the pose information of the input image. Then, this method takes a single-view image for 3D shape representation, which is rendered from a shape under the predicted pose. An extra CNN architecture is employed for 3D model feature extraction. Subsequently, this approach introduces a joint network for cross-modal feature learning, which effectively decreases the gap between modalities. In this procedure, metric learning is exploited to control the cross-modality feature distribution. The performance on the MI3DOR dataset [51] demonstrates the effectiveness of this method. Therefore, we utilized this method in our work for the retrieval of prior knowledge (3D structural information) from the input image.

B. Multimodal Information Learning

In this section, we extract the features of the input single-view image and corresponding 3D shape priors, and then fuse the multimodal information for 3D shape reconstruction. Here, for effective feature learning, we extract domain-specific characteristics from each modality. The retrieved 3D shape priors are selected based on the similarity of 3D structures. Thus, this

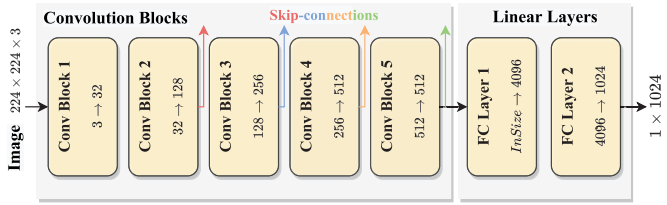


Fig. 2. Architecture of the image encoder. The green, blue, yellow, and red arrows represent the skip connections of the feature maps of the input image.

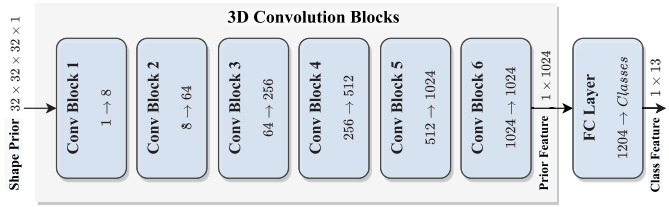


Fig. 3. Architecture of the 3D encoder. We applied the cross-entropy loss function to optimize the parameters of this network.

modality can provide an approximate description of the target 3D structure while still lacking individual details. Considering that the single-view image explicitly describes the target shape from the 2D perspective, we can directly capture the distinguishing characteristics from this modality to compensate for the lack of individual characteristics in the 3D modality.

Based on these analyses, we face three problems: 1) How can the distinguishing individual information be extracted during the feature learning of single-view images? 2) How can the general structure information be captured during the feature learning of the 3D shapes? 3) How can these separate learned information be combined for 3D shape reconstruction? To deal with these problems, we design different network components, which will be introduced in the following subsections.

1) Image Encoder: In this section, we plan to utilize the target image to guide the generation of individual 3D information. Therefore, we need to guarantee that the extracted image feature includes such distinguishing individual information. Here, we redesign the encoder network based on the VGG-16 structure [52]. The details of the image encoder are shown in Fig. 2. We also employ additional branches as skip connections, which can effectively compensate for the distinguishing features of different scales during the reconstruction. The stacked convolutional layers can effectively extract the individual information of different scales. To guide the learning of this distinguishing individual information, the loss function L_{2D} is designed to amplify the individual details of the input image. The loss function is as follows:

$$L_{2D} = -\frac{1}{B}\sum_{i=1}^{B}\log\left(\frac{\exp(f_i f_i^T)}{\sum_{j=1}^{B}\exp(f_i f_j^T)}\right), \tag{1}$$

where i, j is the index of the feature vector, and f denotes the feature vector. f^T represents the image feature's transpose. B is the size of the training data batch. We compute the cross-correlation of each image feature vector in every batch of input samples. Here, each sample can be considered an individual class because the aim of feature learning is to amplify the individual characteristics of the current sample. Thus, we minimize this function to extract more individual information from the extracted image features.

2) Three-Dimensional Shape Prior Encoder: We are inspired by the method in [53] to design the 3D shape prior encoder. This encoder extracts the features of retrieved shape priors. Because the priors provide a general representation of the target 3D structure, the extracted features should include a union of similar structures in the 3D domain. As shown in Fig. 3, we design five 3D convolutional blocks and an extra linear layer to obtain a feature vector of length 1024 from a $32 \times 32 \times 32$ voxel gird. The kernel size of these convolutional layers is 4^3 with padding of 1. The linear layer is used only to adjust the output channels.

To capture more similar structures of the retrieved priors, we introduce the cross-entropy loss function for the parameter learning. The loss function is as follows:

$$L_{3D} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}^{*}\log\left(y_{ic}\right), \tag{2}$$

where N is the number of training samples, and C is the number of classes. y_{ic} represents the predicted probability of class c, and y_{ic}^{*} is the one-hot coding value of the real label. We applied this loss function to capture the common 3D characteristics under each class and to encode these categorical shape structures into the output features.

We also tried to increase the depth of the shape prior encoder. However, additional convolutional layers do not lead to better performance. It not only increases the computation complexity and optimization difficulty, but also causes the information oblivion in shallow layers. However, too few convolution layers fail to capture sufficient structural information for the reconstruction, which is revealed by the wrong reconstruction of categorical parts on the shape. A practical case is that, an input chair with four straight legs was wrongly reconstructed as a board with no legs, which runs counter to the common understanding of the structure of chairs. In the future, we hope to introduce proper residual connections into the deeper prior encoder to further enhance the representational ability of the 3D prior features.

3) Multimodal Information Fusion: The multimodal inputs, the view, and the prior knowledge are encoded to obtain the individual feature vectors and the common feature vectors, respectively. For example, for the input view of a "chair", the view encoder usually focuses on the shape of the armrests, back and seat surface (round, square, tall, short, and hollow, etc.), and the type of legs (straight, curved, or five-star legs, etc.), while the 3D shape encoder focuses on the common characteristics of chairs, for instance, the existence of backs and legs. Obviously, these common features are composed of some individual features, and the individual features provide the structural characteristics of the target object that cannot be ignored. Therefore, we design a multimodal information fusion module, MM-Fusion, based

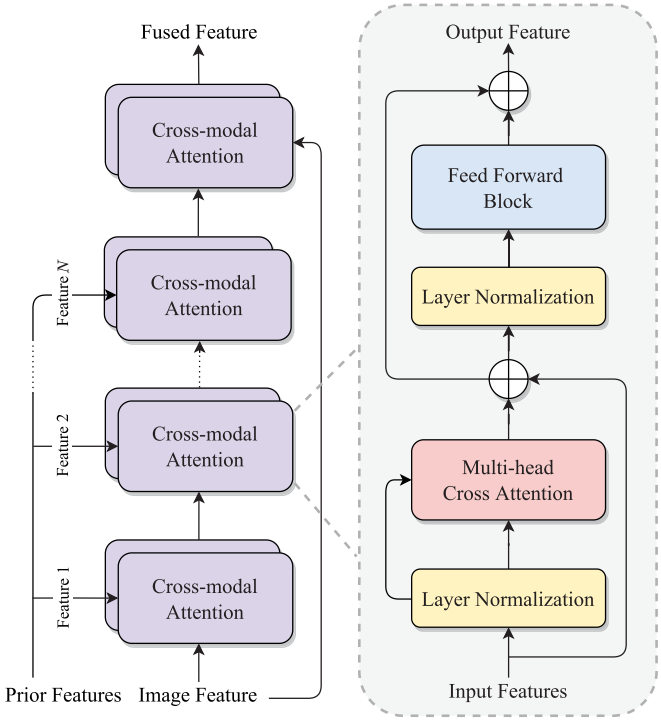


Fig. 4. Detailed structure of the Multimodal information fusion network (MM-Fusion).

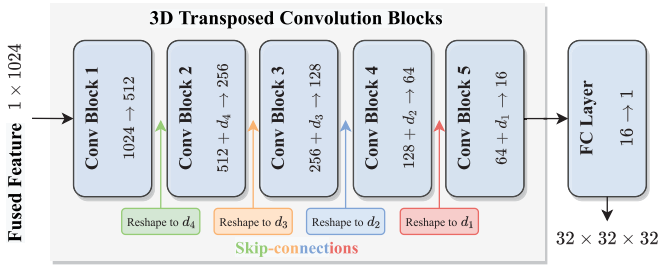


Fig. 5. Architecture of the reconstruction network (Decoder). The input of the decoder is the fused information of the image encoder and the 3D shape encoder's output. Meanwhile, we introduce the skip connections from the image encoder which are indicated by the green, yellow blue and red arrows.

on the attention mechanism in Transformer [54], which focuses on the information interaction between different features and cascades the correlation between them to achieve cross-modal feature fusion.

The structure of the multimodal information fusion module is shown on the left of Fig. 4. The input of the module consists of two parts, image features and prior knowledge feature sets (a collection of features of related 3D prior shapes). As shown in the figure, the fusion module consists of a series of cross-modal attention sub-blocks, and each sub-block can accept two input vectors. They are named as side input term, $\boldsymbol{f}_s$, and bottom input term, $\boldsymbol{f}_b$, according to the input positions in the Fig. 4. Among them, in order to make full use of the prior knowledge, all the subblocks take the output of the previous level one as the $\boldsymbol{f}_b$, except that the first subblock accepts image features as its $\boldsymbol{f}_b$. Similarly, to prevent the network from forgetting the image details, all the subblocks take different 3D prior features as the $\boldsymbol{f}_s$, except the $\boldsymbol{f}_s$ of the last subblock, which accepts the residual connection of image features.

Specifically, each cross-modal attention subblock contains multiple attention units, and the unit structure is shown on the right of Fig. 4. Each unit mainly consists of a multi-head cross attention mechanism and a feed-forward unit. First, the $\boldsymbol{f}_s$ and $\boldsymbol{f}_b$ of each subblock need to be pre-processed by the linear layer. We name the processed results of $\boldsymbol{f}_b$ as the term $\boldsymbol{q}$, while naming the processed results of $\boldsymbol{f}_s$ as terms $\boldsymbol{k}$ and $\boldsymbol{v}$. Then, the cross-modal attention units can be expressed as follows.

$$\boldsymbol{f}_{bn} = \mathrm{LNorm}(\boldsymbol{f}_b), \boldsymbol{f}_{sn} = \mathrm{LNorm}(\boldsymbol{f}_s) \tag{3}$$

$$\boldsymbol{q} = \boldsymbol{f}_{bn}\boldsymbol{W}_Q, \boldsymbol{k} = \boldsymbol{f}_{sn}\boldsymbol{W}_K, \boldsymbol{v} = \boldsymbol{f}_{sn}\boldsymbol{W}_V \tag{4}$$

$$\boldsymbol{z} = \mathrm{MHCA}(\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}) \tag{5}$$

$$\boldsymbol{z}' = \boldsymbol{z} + \boldsymbol{f}_s \tag{6}$$

$$\boldsymbol{z}'' = \mathrm{FFW}\left(\mathrm{LNorm}(\boldsymbol{z}')\right) \tag{7}$$

$$\boldsymbol{f} = \boldsymbol{z}'' + \boldsymbol{z}' \tag{8}$$

where $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$ and $\boldsymbol{W}_V$ are the linear layer weight matrices used for pre-processing. MHCA, LNorm, and FFW represent the multi-head cross attention, layer normalization, and feed forward unit, respectively. The term $\boldsymbol{z}$ represents the multihead attention result, $\boldsymbol{f}$ represents the fusion result of this attention unit, $\boldsymbol{z}'$ and $\boldsymbol{z}''$ is the intermediate result of the unit.

To be specific, layer normalization, LNorm, computes the mean and variance on each input feature and reduces the feature scale to speed up the convergence of training. Feed forward unit, FFW, is a multilayer perceptron with only one hidden layer and uses the GELU (Gaussian Error Linear Unit) function [55] for activation. Multi-headed cross attention can be expressed as follows,

$$\mathrm{CrossAttn}(\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}) = \mathrm{softmax}\left(\boldsymbol{q}\boldsymbol{k}^T/\sqrt{d_\mathrm{k}}\right)\boldsymbol{v} \tag{9}$$

$$\mathrm{Head_h} = \mathrm{CrossAttn}(\boldsymbol{q_h}, \boldsymbol{k_h}, \boldsymbol{v_h}) \tag{10}$$

$$\boldsymbol{z} = \mathrm{Concat}\left(\mathrm{Head}_1, \ldots, \mathrm{Head_H}\right)\boldsymbol{W}_O \tag{11}$$

Here, $\boldsymbol{k}^T$ is the transpose of embedding $\boldsymbol{k}$, d_k represents the length of the vector, $\mathrm{Head_h}$ represents the result of the calculation of the h^{th} attention head (the total number of attention heads is H). The function Concat splices the features on the channel, and the matrix $\boldsymbol{W}_O$ is the weight of the linear layer that processes the splice result.

C. Three-Dimensional Shape Reconstruction

In the shape decoder, we apply a structure that is symmetric with the 3D shape encoder. The decoder needs to transform the previously fused features into 3D volumes. We design a stack of 3D transposed convolutional blocks [56] in the decoder, as shown in Fig. 5. The whole structure receives a combined feature of size 1×1024 and outputs a voxelized 3D shape of size $32 \times 32 \times 32$. These blocks are all composed of a transposed 3D convolution layer followed by a 3D batch normalization layer and a ReLU (Rectified Linear Unit) for optimization, except the

last block, which imposes a sigmoid optimization for the voxel occupation probability.

To obtain sufficient information from the input image, we introduce the image encoder's processing information into the transposed convolutional layers in the decoder using skip connections [57], [58]. These connections are illustrated as the colored arrows in Figs. 2 and 5 (Arrows in the same color means they are the same skip connection). This operation is to compensate for the oblivion of the input information during 3D shape reconstruction. Note that the feature maps from the image encoder is reshaped to match the input size of its skip-connected 3D transposed convolutional layer.

Furthermore, we introduce the refiner network inspired by [59]. It aims to correct the wrongly recovered parts of a reconstructed 3D volume. Following the concept of a 3D encoder-decoder with U-net connections, the refiner preserves the local structure in the fused volume to generate optimized 3D volumes at 32^3 resolution.

1) Optimization: The loss function of the network is defined as the mean value of the voxelwise binary cross-entropy (BCE) between the reconstructed object and the ground truth. More formally, it can be defined as

$$L_o = \frac{1}{S}\sum_{i=1}^{S}\left[gt_i \log\left(p_i\right) + \left(1-gt_i\right)\log\left(1-p_i\right)\right] \tag{12}$$

where S is the number of reconstructed voxel units per shape. gt represents the ground truth shape, and p is the predicted occupancy probability. The decrease in the L_o value demonstrates that the prediction approaches the ground truth.

Overall, the loss function L_{2D} and L_{3D} are used for updating the parameters of the image encoder and shape encoder, respectively. L_o reflects the reconstruction quality and, thus, is employed for updating the parameters of the entire network. Besides, we employ an additional BCE loss, L_r, to train the Shape Refiner. The classic Adam optimizer [60] is used to handle the optimization.

IV. Experiments

In this section, we present extensive experimental evaluations of CPG3D on the ShapeNet [61] and Pix3D datasets [62]. We first describe the datasets and evaluation protocols. Then, the implementation details of the proposed methods are shown briefly. Above all, we report experimental evaluations of the proposed methods against state-of-the-art methods. Based on this, the ablation study and analysis further reveal the inner rationality and effectiveness of the proposed method.

A. Datasets

ShapeNet: ShapeNet [61] is announced as an ongoing effort to establish a richly annotated large-scale dataset of 3D shapes. However, the complete ShapeNet dataset is still not publicly available. For a more convenient use of the dataset, ShapNetCore V1 was released as 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 approximately 51,300 unique 3D models, but no rendered images are provided. Thus, we use the renderings provided by Choy et al. [7]. They rendered every model into 24 views and voxelized the models into 3D voxels while following the original naming and classifying strategies. The 24 randomly rendered views of each 3D model are of size 137×137, and voxelized 3D shapes are of size $32 \times 32 \times 32$. In addition, a uniform colored background is applied to the image during the experiment.

Pix3D: The Pix3D dataset [62] is a large-scale benchmark of diverse image-shape pairs with pixel-level 2D-3D alignment that is specially built for image-based reconstruction tasks. The dataset contains 395 real-world models and 10,069 images of 9 object classes. Different from other previous datasets, Pix3D possesses both real-world images and precise 2D-3D alignment while maintaining the divergence of 3D models. The Pix3D dataset comes with rich information about each image-shape pair: 2D and 3D key points, voxel representation, image mask, and rendering camera intrinsic and extrinsic parameters. Moreover, Pix3D denotes the occlusion and truncation of each model as fields in the comprehensive document.

B. Evaluation Metrics

Before evaluating the reconstruction quality of the proposed method, we binarize the reconstructed probability voxels at a fixed threshold of 0.4 and thus output the standard voxel shapes. For the similarity measure between reconstructed shape and ground truth, we use intersection over union (IoU), which suits volumetric approaches best. More formally,

$$\text{IoU} = \frac{\sum_{i,j,k}\mathbb{I}\left(\hat{y}_{(i,j,k)} > \tau\right) \wedge \mathbb{I}\left(y_{(i,j,k)}\right)}{\sum_{i,j,k}\mathbb{I}\left(\hat{y}_{(i,j,k)} > \tau\right) \vee \mathbb{I}\left(y_{(i,j,k)}\right)} \tag{13}$$

where τ is the threshold of the previously mentioned binarization while $\mathbb{I}$ represents an indication function. The term y and $\hat{y}$ represent the corresponding occupancy of the ground truth and the predicted probability at voxel position (i, j, k), respectively.

For a more robust indicator of the reconstruction performance, Tatarchenko et al. [63] proposed an F-score that calculates the harmonic mean between precision and recall. In this case, precision indicates the accuracy of the reconstruction, and recall indicates the completeness of the reconstruction. More formally,

$$\text{F-score}(d) = \frac{2P(d)R(d)}{P(d)+R(d)} \tag{14}$$

where $P(d)$ and $R(d)$ represent the precision and recall at a fixed distance threshold d, respectively. Specifically,

$$P(d) = \frac{1}{n_{\mathcal{R}}}\sum_{r\in\mathcal{R}}\left[\min_{g\in\mathcal{G}}\|g-r\| < d\right] \tag{15}$$

$$R(d) = \frac{1}{n_{\mathcal{G}}}\sum_{g\in\mathcal{G}}\left[\min_{r\in\mathcal{R}}\|g-r\| < d\right] \tag{16}$$

where $\mathcal{R}$ and $\mathcal{G}$ denote the reconstructed and ground truth point clouds, respectively. The term n with subscripts $\mathcal{R}$ and $\mathcal{G}$ represents the scale of the corresponding point clouds. The F-score works on the point clouds and concerns the distance between the points and surfaces. Therefore, for volumetric approaches,

TABLE I
COMPARISON OF SINGLE-VIEW 3D OBJECT RECONSTRUCTION ON SHAPENET AT 32^3 RESOLUTION. WE REPORT THE MEAN IOU PER CATEGORY

Category	3D-R2N2 [7]	OGN [26]	Pixel2Mesh [29]	AttSets [39]	Pix2Vox++ [40]	DASI [13]	Mem3D [11]	CPG3D
Airplane	0.513	0.587	0.508	0.594	0.674	0.701	0.767	**0.772**
Bench	0.421	0.481	0.379	0.552	0.608	0.625	0.651	**0.659**
Cabinet	0.716	0.729	0.732	0.783	0.799	0.798	0.840	**0.875**
Car	0.798	0.828	0.670	0.844	0.858	0.861	0.877	**0.894**
Chair	0.466	0.483	0.484	0.559	0.581	0.578	0.712	**0.724**
Display	0.468	0.502	0.582	0.565	0.548	0.552	0.631	**0.648**
Lamp	0.381	0.398	0.399	0.445	0.457	0.470	0.535	**0.544**
Speaker	0.662	0.637	0.672	0.721	0.721	0.723	0.778	**0.791**
Rifle	0.544	0.593	0.468	0.601	0.617	0.652	0.746	**0.758**
Sofa	0.628	0.646	0.622	0.703	0.725	0.723	0.753	**0.770**
Table	0.513	0.536	0.536	0.590	0.620	0.614	0.685	**0.692**
Telephone	0.661	0.702	0.762	0.743	0.809	0.801	0.823	**0.836**
Watercraft	0.590	0.632	0.471	0.601	0.603	0.622	0.684	**0.695**
Overall	0.560	0.596	0.552	0.642	0.670	0.676	0.729	**0.743**

The best result for each category is highlighted in bold.

TABLE II
COMPARISON OF SINGLE-VIEW 3D OBJECT RECONSTRUCTION ON SHAPENET AT 32^3 RESOLUTION. WE REPORT THE F-SCORE@1% PER CATEGORY

Category	3D-R2N2 [7]	OGN [26]	Pixel2Mesh [29]	AttSets [39]	Pix2Vox++ [40]	DASI [13]	Mem3D [11]	CPG3D
Airplane	0.412	0.487	0.376	0.489	0.583	0.604	0.671	**0.679**
Bench	0.345	0.364	0.313	0.406	0.478	0.484	0.525	**0.531**
Cabinet	0.327	0.316	0.450	0.367	0.408	0.431	0.517	**0.536**
Car	0.481	0.514	0.486	0.497	0.564	0.574	0.590	**0.598**
Chair	0.238	0.226	0.386	0.334	0.309	0.296	0.503	**0.511**
Display	0.227	0.215	0.319	0.310	0.296	0.299	0.498	**0.507**
Lamp	0.267	0.249	0.219	0.315	0.315	0.336	0.403	**0.414**
Speaker	0.231	0.225	0.190	0.211	0.152	0.294	0.262	**0.233**
Rifle	0.521	0.541	0.340	0.524	0.574	0.604	0.626	**0.632**
Sofa	0.274	0.290	0.343	0.334	0.377	0.393	0.434	**0.445**
Table	0.340	0.352	0.502	0.419	0.406	0.392	0.569	**0.571**
Telephone	0.504	0.528	0.485	0.469	0.633	0.638	0.674	**0.683**
Watercraft	0.305	0.328	0.266	0.315	0.390	0.425	0.461	**0.470**
Overall	0.351	0.368	0.398	0.395	0.436	0.443	0.517	**0.524**

The best result for each category is highlighted in bold.

we utilize the marching cubes algorithm [64] to construct the object surface and then sample 8,192 points from this surface for the F-score calculation. For the surface-based methods without explicit points, we also use the marching cube algorithm to construct the surface and then sample 8,192 points to compute the F-Score.

C. Implementation Details

Our network inputs are set to RGB images on the scale of 224×224, and the retrieved voxels are all in the size of $32 \times 32 \times 32$. Note that the original volumetric shapes in the Pix3D dataset are of size $128 \times 128 \times 128$. Thus, we downsampled the input voxel shapes into 32^3 to align with the input channel of the shape prior encoder. The volumetric output is identical to the input voxels in size. For single-view images, we use pre-trained VGG-16 [52]. The retrieved voxels are encoded by the stacked 3D convolutional layers. To effectively utilize the multimodal features, we set each of the cross-modal attention blocks as two attention units with six heads. On the decoder, we exploit five transposed 3D convolutional layers to reconstruct the volumes. During the training procedure, we implement our network in PyTorch with a data batch size of 29 on a single Nvidia GTX 1080Ti GPU, and the unfull batch is discarded. In addition, we use an Adam optimizer [60] with a β_1 of 0.9 and β_2 of 0.999. The initial learning rate is set to 10^{-5} and decays to a half after 150 epochs.

D. Single-View Reconstructions

1) Evaluation on ShapeNet: To demonstrate the performance of our approach, we compare our method with several SOTA methods on the ShapeNet dataset. To ensure a fair comparison, all methods are evaluated under the same input images for all experiments. The final experimental results are listed in Tables I and II. Higher scores of IoU and F-Score@1% indicate better reconstruction quality.

Fig. 6 shows several reconstruction examples on the ShapeNet testing set. In these examples, benefiting from the retrieved shape priors, CPG3D maintains the fundamental structures of the target shape, such as the legs of the chair and the spoiler wing of the sports car. Meanwhile, CPG3D also captures more details of the input images, for instance, the shapes of sofa armrests, holes in stool legs, and wheels of the car. Compared with the implicit surface reconstruction method DISN, our method recovers better details of the target shapes, but DISN always reconstructs smoother surfaces, which is the inborn advantage of the implicit surfaces.

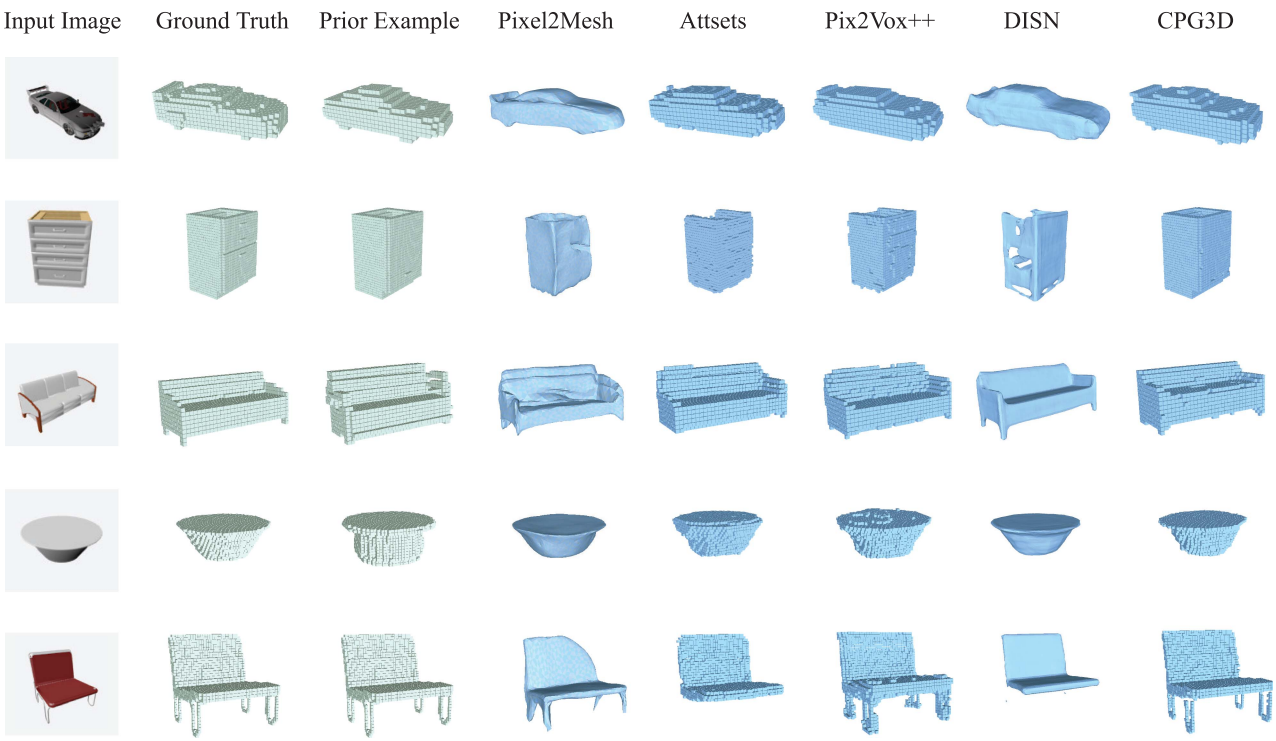


Fig. 6. Examples of single-view 3D object reconstruction on the ShapeNet dataset.

TABLE III
COMPARISON OF SINGLE-VIEW 3D OBJECT RECONSTRUCTION ON PIX3D AT 32^3 RESOLUTION USING MEAN IOU AND F-SCORE@1%

Method	IoU	F-Score@1%
3D-R2N2 [7]	0.136	0.018
3D-VAE-GAN [65]	0.171	-
MarrNet [66]	0.231	0.026
DRC [67]	0.265	0.038
ShapeHD [68]	0.284	0.046
DAREC [69]	0.241	-
Pix3D [62]	0.267	0.041
Pix2Vox++ [40]	0.288	0.068
DASI [13]	0.310	-
FroDo [70]	0.325	-
Mem3D [11]	0.387	0.143
CPG3D	**0.391**	**0.152**

The best result is highlighted in bold.

TABLE IV
COMPARISON OF SINGLE-VIEW 3D OBJECT RECONSTRUCTION ON SHAPENET AT 64^3 AND 128^3 RESOLUTION. WE REPORT THE MEAN IOU AND F-SCORE@1%

Method	Resolution-64^3		Resolution-128^3	
	IoU	F-Score@1%	IoU	F-Score@1%
OGN [26]	0.771	0.361	0.782	0.390
Matryoshka [71]	0.784	0.380	0.794	0.426
Pix2Vox++ [40]	0.803	0.418	0.826	0.475
CPG3D	**0.839**	**0.455**	**0.860**	**0.521**

The best result of each metric is highlighted in bold.

2) Evaluation on Pix3D: We also evaluated the performance on the Pix3D dataset. Pix3D aligned the 3D models precisely with the images and marked the shapes of occlusion and truncation. Thus, the results on the clear images can verify the performance on single-view reconstruction, and the results on the disturbed images demonstrate the positive effects of the retrieved shape priors. The overall quantitative results are shown in Table III. CPG3D achieves the best performance compared with previous SOTA methods. We visualized some reconstructed examples and corresponding retrieved priors in Fig. 7. The quantitative results show that the introduction of shape priors can also improve the reconstruction qualities of real-world images. In addition, we reconstruct both the occluded images and unoccluded images in Figs. 8 and 9, respectively. The visualized results on unoccluded shapes show that CPG3D outperforms previous methods and recovers more details of the target shapes. On the other hand, the visualized results on occluded shapes obviously outperform the previous methods. CPG3D correctly recovers the occluded parts in the input images, while other methods are confused by these disturbances. These evaluated results indicate that CPG3D generates more compelling shapes on images with complicated backgrounds and disturbances such as occlusion.

3) Higher-Resolution 3D Object Reconstruction: Shapes in low resolution have better portability due to their lightness in size. However, this property also signifies their lack of high-frequency shape details. Therefore, unlike the commonly researched resolution of 32^3, we fit our method to reconstruct shapes in higher resolution to evaluate the performance of our 3D-prior-based method in handling more complex details. Following the experiment settings in OGN [26], we report the mean IoU and F-Score@1% in the resolution of 64^3 and 128^3 in Table IV. The experimental results show that our method outperforms Pix2Vox, Matryoshka [71], and OGN by a large margin.

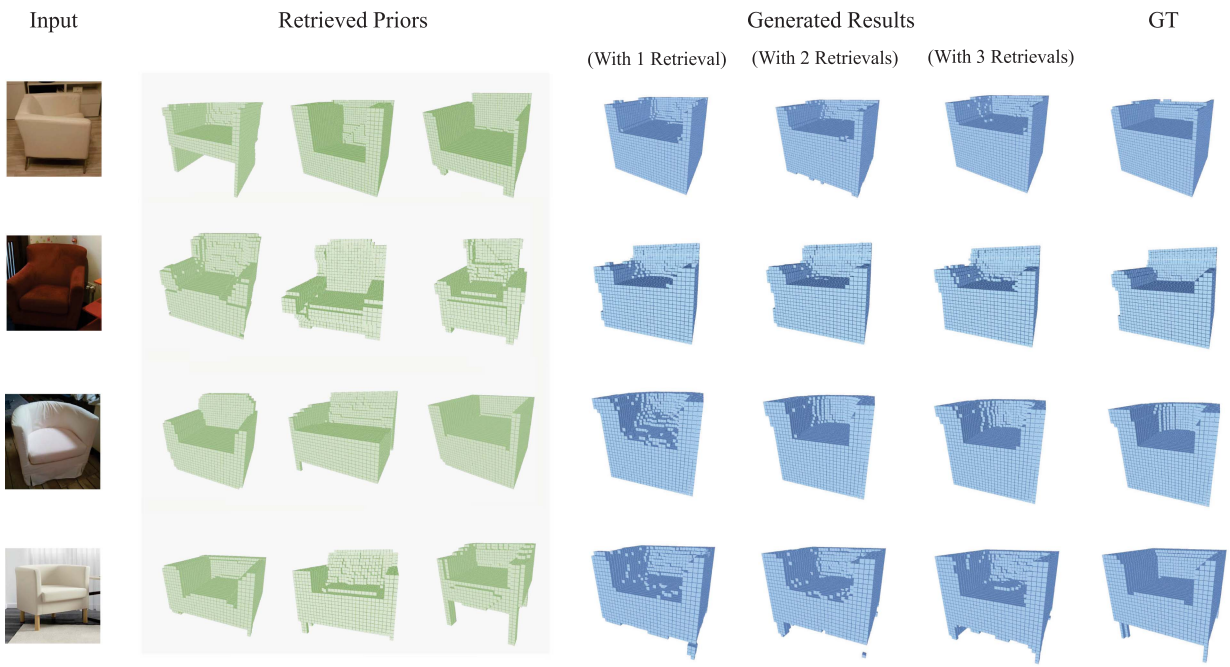


Fig. 7. Single-view 3D object reconstructions and related retrieved priors on the Pix3D dataset.

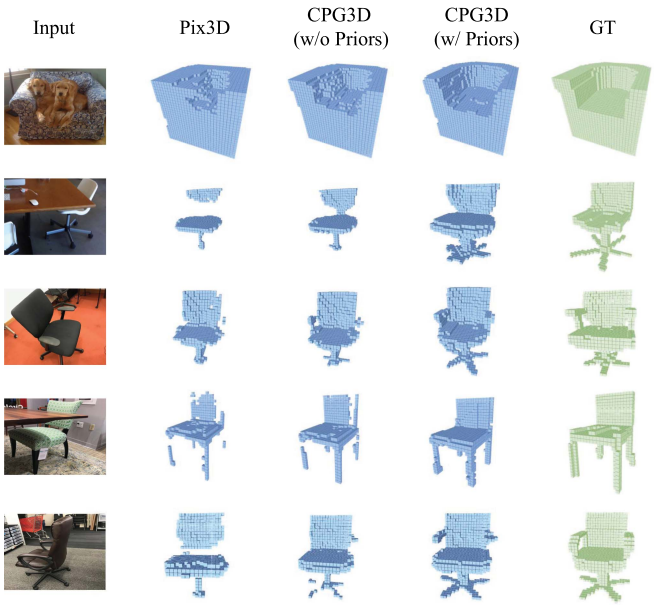


Fig. 8. Reconstructions of occluded samples on the Pix3D dataset.

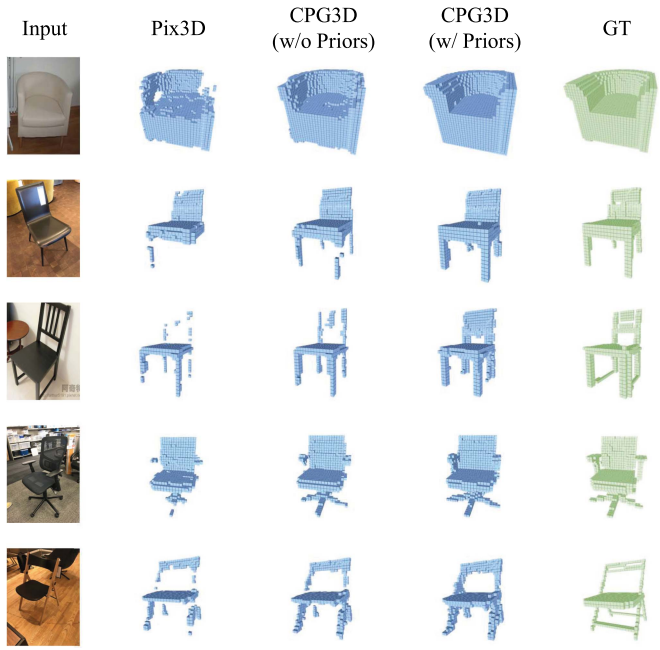


Fig. 9. Reconstructions of unoccluded samples on the Pix3D dataset.

Furthermore, to illustrate the reconstructing ability, we visualized several examples in Fig. 10, as shown in which, our method is more capable of capturing high-resolution details than the previous method, such as the hubs, pedals, police lights, and bumpers.

E. Ablation Study

1) Network Structure Ablation: There are four modules in this network: image encoder, 3D model encoder, feature fusion module, and shape reconstruction module. We performed ablation experiments on these modules to demonstrate the rationality and superiority of our design. The quantitative results on the ShapeNet dataset are listed in Table V.

a) Ablation on image encoder: Our shape reconstruction occurs under the guidance of shape priors. Nevertheless, the retrieved shape priors focus on the recovery of structurally similar areas, which leads to unsatisfactory reconstructions. Relevant experimental results are in Table V.

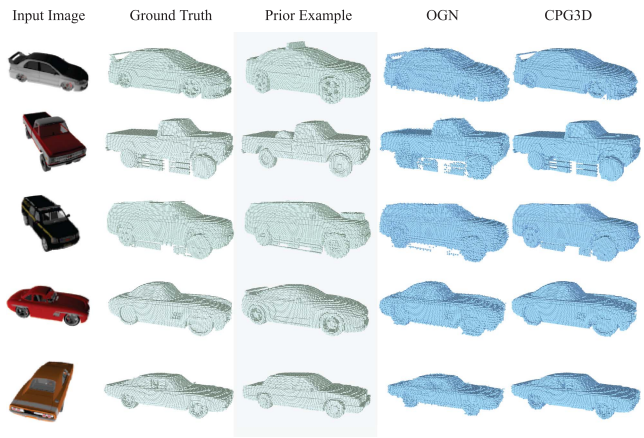


Fig. 10. Single-view reconstructions on ShapeNet-car dataset at 128^3 resolution.

TABLE V
ABLATION STUDY ON THE NETWORK STRUCTURE ON SHAPENET DATASET

Structure	- Batchsize	IoU	F-Score@1%
CPG3D	- 29	**0.743**	**0.524**
w/o Refiner	- 35	0.723	0.498
	- 29	0.722	0.498
	- 8	0.718	0.475
w/o Shape Encoder	- 35	0.702	0.453
	- 29	0.698	0.451
	- 8	0.676	0.432
w/o Image Encoder	- 35	0.667	0.428
	- 29	0.664	0.427
	- 8	0.660	0.422

b) Ablation on shape encoder: To demonstrate the effectiveness of shape priors, we performed an ablation study on the shape encoder, i.e., reconstruction from input images. Although the quantitative results are comparable to those of some previous methods, there is still room for improvement when compared with our prior-guided pipeline.

c) Ablation on refiner: Our reconstructions from the shape decoder are further optimized in a U-net [58] manner, which is defined as refiner in the shape reconstruction module. Here, we pruned refiner to test the optimization effects. The quantitative results in Table V demonstrate the important role of Refiner in improving reconstruction quality.

d) Different batch sizes: Our baselines are divided into w/o Refiner, w/o Shape Encoder, and w/o Image Encoder. We conducted experiments of these baselines under different batch sizes. Under the batch size of 29, the entire pipeline achieved the best performance. According to the experimental results in Table V, the performance of baselines improves with increasing batch size, but the improvement trend gradually stabilizes. The results of the baselines do not exceed the entire pipeline's performance. In conclusion, the batch size indeed influences the performance of the baselines. However, the effect is limited. The baseline methods, which lack relevant modules, cannot get the best result beyond the entire pipeline of CPG3D. It indirectly proves the necessity and rationality of each module in our approach.

TABLE VI
EFFECTS OF DIFFERENT LOSS FUNCTIONS ON SHAPENET DATASET. (NOTE THAT WE USE "+" TO REPRESENT "AND" BUT NOT MATHEMATICAL PLUS OPERATION.)

Loss	IoU	F-Score@1%
L_{BCEs}	0.703	0.468
$L_{BCEs} + L_{2D(CE)}$	0.708	0.472
$L_{BCEs} + L_{3D}$	0.716	0.482
$L_{BCEs} + L_{2D(CE)} + L_{3D}$	0.720	0.488
$L_{BCEs} + L_{2D}$	0.731	0.512
$L_{BCEs} + L_{2D} + L_{3D}$	**0.743**	**0.524**

2) Loss Functions: We utilized different loss functions for each module in our network: (1) L_{2D}: Optimizing the image encoder by enhancing the differences between samples in a batch (note that $L_{2D(CE)}$ is a cross-entropy variant of this function); (2) L_{3D}: a cross-entropy loss function concentrating on the categorical information of 3D priors; (3) L_o and L_r: two binary cross-entropy loss functions comparing the 3D occupancy between the predicted shape and corresponding ground truth. In this part, we validate the effects under different loss functions. Related experimental results are listed in Table VI. Note that BCE loss is indispensable for shape reconstruction, and thus our experiments reveal the effects of other losses based on L_o and L_r. For simplicity, we note $L_o + L_r$ as L_{BCEs}.

It is obvious that images lack the details of unseen parts, and the shape priors struggle to recover object-specific details. This situation requires our network to capture more individual details from input images and more generic structures from combined shape priors. The commonly used loss function $L_{2D(CE)}$ focuses on category-level features, while our L_{2D} emphasizes the importance of object-divergent features. The improvement from $L_{2D(CE)} + L_{BCEs}$ to $L_{2D} + L_{BCEs}$ demonstrates that the object-divergent information indeed benefits the reconstruction. Consistent with the above conclusions, $L_{2D} + L_{BCEs}$ outperforms $L_{3D} + L_{BCEs}$ and even $L_{2D(CE)} + L_{3D} + L_{BCEs}$ due to the incorporation of individual features. The loss term $L_{2D} + L_{3D} + L_{BCEs}$ takes both individual details and generic structures into consideration and undoubtedly achieves the best performance.

3) Skip Connections: The deep neural network is difficult to optimize due to overfitting, unstable random initialization, and gradient problems. Inspired by the designs of ResNet [72] and U-Net [73], we introduced skip connections between the image encoder and the reconstruction decoder. Although this design has widely proven to be effective for network convergence, the connection orders still remain to be further studied. To weigh the improvement benefited from skip connections against the computational consumption caused by the same scheme, we evaluated our network on different orders of such connections, and the quantitative results are listed in Table VII. Here, following the structure in Fig. 1, Skc represents the skip connection, and the indexes ascend from shallow to deep in the image encoder. Experimental results show that the relative high-resolution features

TABLE VII
EXPERIMENTS ON DIFFERENT SKIP-CONNECTION ORDERS ON SHAPENET DATASET

Skip-connection(s)	IoU↑	VRAM Usage (MB)↓
$Skc4$	0.668	11031
$Skc3$	0.673	10864
$Skc2$	0.678	10680
$Skc1$	0.683	**10653**
$Skc2 + Skc3 + Skc4$	0.716	11124
$Skc1 + Skc3 + Skc4$	0.721	11097
$Skc1 + Skc2 + Skc4$	0.725	10943
$Skc1 + Skc2 + Skc3$	0.728	10869
$Skc1 + Skc2 + Skc3 + Skc4$	**0.743**	11151

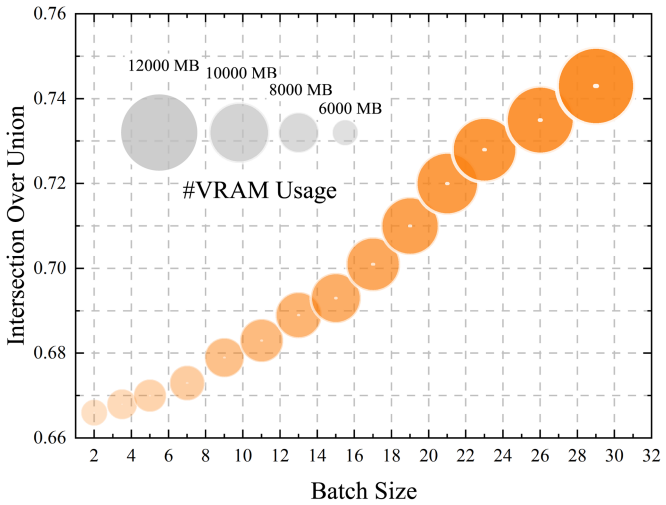


Fig. 11. IoU score and VRAM usage of different batch sizes on ShapeNet dataset.

from shallow layers can compensate for the information in deep layers in the decoder and thus produce better reconstructions.

4) Influence of Different Batch Sizes: We optimized image loss L_{2D} to capture the divergent information between input images. Because L_{2D} is computed within a batch, we evaluated the network on different batch sizes. As illustrated in Fig. 11, the size of the bubbles is proportional to the VRAM usage. A larger batch size tends to capture more divergent information while the VRAM consumption increases simultaneously. We perform the experiment on a single NVIDIA GTX 1080Ti, which supports a maximum batch size of 29, and we will explore the performance on a larger batch size in future work.

F. Analysis

1) Influence of Disturbances on Retrieved Priors:

a) Positive disturbances: Our cross attention module is sensitive to the size of the input feature sequence, which is experimented as a positive disturbance on the retrieved priors. In this part, we trained our network on different numbers of retrieved priors and used IoU as the criterion. The results are visualized as the orange-colored histogram in Fig. 12. Considering that the function L_{3D} leads to more general representations of similar 3D structures, our newly increased shape priors can enrich these representations and thus positively influence our reconstruction.

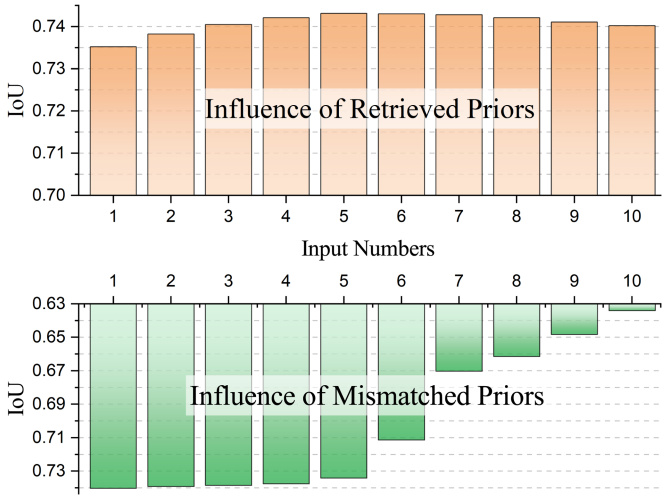


Fig. 12. Influence of Disturbances on Retrieved Priors on ShapeNet dataset.

Nevertheless, this positive impact is not unlimited due to the redundancy of similar information. Our experiment shows that the increase is almost halted when there are more than five input priors.

b) Negative disturbance: For some samples with unique structures, the increasing number of retrieved priors inevitably introduces priors of low retrieval confidence. Here, we artificially introduce specific numbers of unmatched priors to evaluate the robustness of the network. The qualitative results (green-colored histogram in Fig. 12) show that our network can maintain the performance when the number of mismatched priors is less than 3. Note that the extreme situation of input ten mismatched priors still reached the IoU of 0.634, which also verified the robustness of our network on this disturbance.

2) Comparison With Best Retrieval Baseline: As mentioned in [63], encoder-decoder structured networks easily fall into the *recognition* [63] problem, which merely pertains to the whole object and only performs the retrieval on the dataset. To test the upper bound of the retrieval methods on ShapeNet, we evaluated the Oracal Nearest Neighbour (Oracal NN) baseline, which, based on our shape encoder, finds the most similar shape from the training set for shapes in the test set and calculates IoU between the ground truth and the retrieved shape. The Oracle NN baseline demonstrates the best performance of the retrieval methods, which, on the other hand, confirms that our method is not a pure *recognition* network. The experimental results are listed in Table VIII.

3) Analysis of Feature Fusion Methods: To handle the fusion of multimodal information, we modified the cross attention mechanism from Transformer [54] and designed the MM-Fusion module. To verify the effect of this design, we compare it with some classic information fusion strategies. The related experimental results in Table IX show that MM-Fusion outperforms other methods. This structure can effectively reduce the redundant information and thus guarantee the effectiveness of the fused features.

4) Analysis of Feature Distribution: In this article, we utilize the retrieved 3D priors to supplement structural information

TABLE VIII
COMPARISON WITH ORACLE NN ON SHAPENET DATASET

Category	Airplane	Bench	Cabinet	Car	Chair	Display	Lamp
CPG3D	**0.772**	**0.659**	**0.875**	**0.894**	**0.724**	**0.648**	**0.544**
Mem3D	0.767	0.651	0.840	0.877	0.712	0.631	0.535
OracleNN	0.540	0.341	0.576	0.778	0.291	0.412	0.218

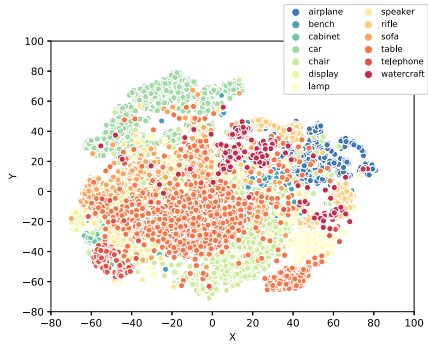
Category	Speaker	Rifle	Sofa	Table	Telephone	Watercraft	Overall
CPG3D	**0.791**	**0.758**	**0.770**	**0.692**	**0.836**	**0.695**	**0.743**
Mem3D	0.778	0.746	0.753	0.685	0.823	0.684	0.729
OracleNN	0.498	0.484	0.499	0.293	0.805	0.442	0.458

TABLE IX
EFFECTS OF DIFFERENT FUSION METHODS ON SHAPENET DATASET

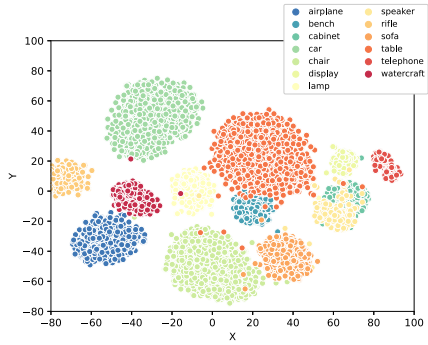
Method	IoU	F-Score@1%
Addition	0.436	0.205
Straight-Connect	0.602	0.351
Average-Pooling	0.622	0.377
Max-Pooling	0.671	0.429
MM-Fusion	**0.743**	**0.524**

lost due to occlusion and background noise. Assume that the reconstruction target is a chair with rectangular arms and four straight legs. The retrieved priors capture this structural information while they are different in details, such as the thickness of the legs or the size of the arms. In other words, these priors are clustered according to the general 3D structures. However, the goal of 3D reconstruction is to recover the exact object from the input image but not to retrieve a similar 3D structure. Thus, we apply a correlation loss L_{2D} to guide the image feature learning to capture more characteristic features. Here, we apply the tSNE [74], [75] toolbox to visualize the distribution of extracted image features and prior features in Fig. 13. As illustrated in the figure, the shape priors clustered tighter according to the class to which they belong, but in some special categories, the network struggles to distinguish features such as the 'Cabinet' and 'Speaker' because their ShapeNet ground truth volumes are similar in size and also share some common characters in 3D appearance. On the other hand, the image features are loosely distributed because they are more concerned with individual-level discrimination. This observation validated the rationality of our motivation as well as the effectiveness of the network design.

5) Analysis of Generalization: To test the generalization of CPG3D, we processed ShapeNet categories beyond the 13-category data used in previous experiments in IV-D and performed evaluations on CPG3D. Note that our priors are also fine-tuned on novel categories. Some experimental results are visualized in Fig. 14. As shown in the figure, CPG3D captures the general structures of target shapes, but there is still a gap between the reconstructed details and ground truths. Our method basically recovers the overall structure information. Furthermore, comparing reconstructed shapes with and without priors, CPG3D indeed learns essential information for better reconstructions from shape priors. This demonstrates the necessity of introducing prior information.



(a)



(b)

Fig. 13. (a) Distribution of features extracted by the image encoder on ShapeNet dataset. (b) Distribution of features extracted by the shape prior encoder.

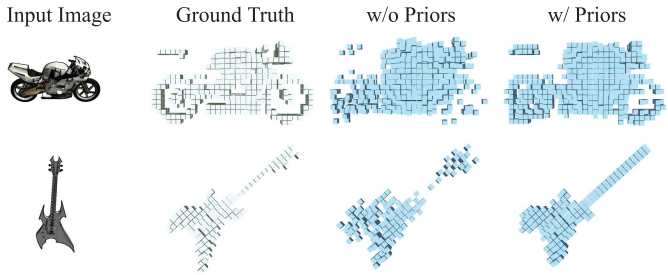


Fig. 14. Reconstrucions on unseen categories on ShapeNet dataset.

V. CONCLUSION

In this article, we proposed a novel approach for reconstructing high-quality 3D objects based on a single-view image and its related 3D shape prior knowledge. We creatively utilized the image to retrieve similar 3D shapes as prior knowledge. These retrieved priors can provide effective geometric information for 3D reconstruction. Meanwhile, we proposed a novel cross-modal integration network for the joint learning of views and priors. Subsequently, we reconstruct the object volume from integrated features and introduced skip connections to optimize the reconstruction. Experimental results demonstrated that our approach significantly improves the reconstruction quality and performs favorably against state-of-the-art methods on the ShapeNet and Pix3D datasets.

REFERENCES

[1] X. Wang, Y. Guo, Z. Yang, and J. Zhang, "Prior-guided multi-view 3D head reconstruction," *IEEE Trans. Multimedia*, vol. 24, pp. 4028–4040, 2022.

[2] X. Fan et al., "Dual neural networks coupling data regression with explicit priors for monocular 3D face reconstruction," *IEEE Trans. Multimedia*, vol. 23, pp. 1252–1263, 2021.

[3] X. Tu et al., "3D face reconstruction from a single image assisted by 2D face images in the wild," *IEEE Trans. Multimedia*, vol. 23, pp. 1160–1172, 2021.

[4] P. Hu, E. S. Ho, and A. Munteanu, "3DBodyNet: Fast reconstruction of 3D animatable human body shape from a single commodity depth camera," *IEEE Trans. Multimedia*, vol. 24, pp. 2139–2149, 2022.

[5] C. Yan et al., "3D room layout estimation from a single RGB image," *IEEE Trans. Multimedia*, vol. 22, no. 11, pp. 3014–3024, Nov. 2020.

[6] Z. Zhou, F. Shi, J. Xiao, and W. Wu, "Non-rigid structure-from-motion on degenerate deformations with low-rank shape deformation model," *IEEE Trans. Multimedia*, vol. 17, no. 2, pp. 171–185, Feb. 2015.

[7] 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. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 628–644.

[8] 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.*, 2017, pp. 605–613.

[9] C. Wen, Y. Zhang, Z. Li, and Y. Fu, "Pixel2Mesh++: Multi-view 3D mesh generation via deformation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1042–1051.

[10] C. Liu, D. Kong, S. Wang, J. Li, and B. Yin, "DLGAN: Depth-preserving latent generative adversarial network for 3D reconstruction," *IEEE Trans. Multimedia*, vol. 23, pp. 2843–2856, 2021.

[11] S. Yang, M. Xu, H. Xie, S. Perry, and J. Xia, "Single-view 3D object reconstruction from shape priors in memory," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3152–3161.

[12] A. Siddique and S. Lee, "Sym3DNet: Symmetric 3D prior network for single-view 3D reconstruction," *Sensors*, vol. 22, no. 2, 2022, Art. no. 518.

[13] J. Gao, D. Kong, S. Wang, J. Li, and B. Yin, "DASI: Learning domain adaptive shape impression for 3D object reconstruction," *IEEE Trans. Multimedia*, early access, Jul. 07, 2022, doi: 10.1109/TMM.2022.3189247.

[14] J. J. Atick, P. A. Griffin, and A. N. Redlich, "Statistical approach to shape from shading: Reconstruction of three-dimensional face surfaces from single two-dimensional images," *Neural Comput.*, vol. 8, no. 6, pp. 1321–1340, 1996.

[15] R. Zhang, P. Tsai, J. E. Cryer, and M. Shah, "Shape from shading: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 8, pp. 690–706, Aug. 1999.

[16] R. Dovgard and R. Basri, "Statistical symmetric shape from shading for 3D structure recovery of faces," in *Proc. 8th Eur. Conf. Comput. Vis.*, 2004, pp. 99–113.

[17] G. Riegler, A. O. Ulusoy, H. Bischof, and A. Geiger, "OctNetFusion: Learning depth fusion from data," in *Proc. IEEE Int. Conf. 3D Vis.*, 2017, pp. 57–66.

[18] L. Guan, J. S. Franco, and M. Pollefeys, "3D occlusion inference from Silhouette cues," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[19] J. Aliomonos and M. J. Swain, "Shape from texture," in *Proc. 9th Int. Joint Conf. Artif. Intell.*, 1985, pp. 926–931.

[20] A. M. Loh and R. I. Hartley, "Shape from non-homogeneous, non-stationary, anisotropic, perspective texture," in *Proc. Brit. Mach. Vis. Conf.*, 2005, pp. 69–78.

[21] P. Parodi and G. Piccioli, "3D shape reconstruction by using vanishing points," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 2, pp. 211–217, Feb. 1996.

[22] D. Meagher, "Geometric modeling using octree encoding," *Comput. Graph. Image Process.*, vol. 19, no. 2, pp. 129–147, 1982.

[23] A. Miller, V. Jain, and J. L. Mundy, "Real-time rendering and dynamic updating of 3-D volumetric data," in *Proc. 4th Workshop Gen. Purpose Process. Graph. Process. Units*, 2011, pp. 1–8.

[24] G. Riegler, A. O. Ulusoy, and A. Geiger, "OctNet: Learning deep 3D representations at high resolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6620–6629.

[25] C. Hane, S. Tulsiani, and J. Malik, "Hierarchical surface prediction for 3D object reconstruction," in *Proc. IEEE Int. Conf. 3D Vis.*, 2017, pp. 412–420.

[26] 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.*, 2017, pp. 2107–2115.

[27] P. Wang, C. Sun, Y. Liu, and X. Tong, "Adaptive O-CNN: A patch-based deep representation of 3D shapes," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 1–11, 2018.

[28] H. Kato, Y. Ushiku, and T. Harada, "Neural 3D mesh renderer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3907–3916.

[29] N. Wang et al., "Pixel2mesh: Generating 3D mesh models from single RGB images," in *Proc. 15th Eur. Conf. Comput. Vis.*, 2018, pp. 55–71.

[30] M. Vakalopoulou et al., "AtlasNet: Multi-atlas non-linear deep networks for medical image segmentation," in *Proc. 21st Med. Image Comput. Comput. Assist. Interv.*, 2018, pp. 658–666.

[31] J. K. Pontes et al., "Image2Mesh: A learning framework for single image 3D reconstruction," in *Proc. 14th Asian Conf. Comput. Vis.*, 2018, pp. 365–381.

[32] J. Pan, X. Han, W. Chen, J. Tang, and K. Jia, "Deep mesh reconstruction from single RGB images via topology modification networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9963–9972.

[33] C. Lv, W. Lin, and B. Zhao, "Voxel structure-based mesh reconstruction from a 3D point cloud," *IEEE Trans. Multimedia*, vol. 24, pp. 1815–1829, 2022.

[34] L. M. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3D reconstruction in function space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4460–4470.

[35] Z. Chen and H. Zhang, "Learning implicit fields for generative shape modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5939–5948.

[36] M. Michalkiewicz, J. K. Pontes, D. Jack, M. Baktashmotlagh, and A. Eriksson, "Deep level sets: Implicit surface representations for 3D shape inference," 2019, *arXiv:1901.06802*.

[37] Z. Chen and H. Zhang, "Learning implicit fields for generative shape modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5939–5948.

[38] Q. Xu, W. Wang, D. Ceylan, R. Mech, and U. Neumann, "DISN: Deep implicit surface network for high-quality single-view 3D reconstruction," in *Proc. Adv. Neural Inf. Process. Syst. 32: Annu. Conf. Neural Inf. Process. Syst.*, 2019, pp. 490–500.

[39] B. Yang, S. Wang, A. Markham, and N. Trigoni, "Robust attentional aggregation of deep feature sets for multi-view 3D reconstruction," *Int. J. Comput. Vis.*, vol. 128, no. 1, pp. 53–73, 2020.

[40] H. Xie, H. Yao, S. Zhang, S. Zhou, and W. Sun, "Pix2Vox : Multi-scale context-aware 3D object reconstruction from single and multiple images," *Int. J. Comput. Vis.*, vol. 128, no. 12, pp. 2919–2935, 2020.

[41] J. Weston, S. Chopra, and A. Bordes, "Memory networks," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.

[42] A. H. Miller et al., "Key-value memory networks for directly reading documents," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 1400–1409.

[43] Y. Zhu et al., "Vision-dialog navigation by exploring cross-modal memory," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10727–10736.

[44] W. Wang et al., "Memory-based network for scene graph with unbalanced relations," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 2400–2408.

[45] M. Aubry, D. Maturana, A. A. Efros, B. C. Russell, and J. Sivic, "Seeing 3D chairs: Exemplar part-based 2D-3D alignment using a large dataset of CAD models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3762–3769.

[46] H. Su, S. Maji, E. Kalogerakis, and E. G. Learned-Miller, "Multi-view convolutional neural networks for 3D shape recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 945–953.

[47] B. Shi, S. Bai, Z. Zhou, and X. Bai, "DeepPano: Deep panoramic representation for 3-D shape recognition," *IEEE Signal Process. Lett.*, vol. 22, no. 12, pp. 2339–2343, Dec. 2015.

[48] T. Lee et al., "Cross-domain image-based 3D shape retrieval by view sequence learning," in *Proc. Int. Conf. 3D Vis.*, 2018, pp. 258–266.

[49] P. Mu, S. Zhang, Y. Zhang, X. Ye, and X. Pan, "Image-based 3D model retrieval using manifold learning," *Front. Inf. Technol. Electron. Eng.*, vol. 19, no. 11, pp. 1397–1408, 2018.

[50] W. Nie, Y. Zhao, J. Nie, A. A. Liu, and S. Zhao, "CLN: Cross-domain learning network for 2D image-based 3D shape retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 992–1005, Mar. 2022.

[51] H. Zhou, A. Liu, and W. Nie, "Dual-level embedding alignment network for 2D image-based 3D object retrieval," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1667–1675.

[52] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.

[53] D. Maturana and S. Scherer, "VoxNet: A 3D convolutional neural network for real-time object recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2015, pp. 922–928.
[54] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.: Annu. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[55] D. Hendrycks and K. Gimpel, "Bridging nonlinearities and stochastic regularizers with Gaussian error linear units," 2016, *arXiv:1606.08415*.
[56] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2528–2535.
[57] A. Dosovitskiy et al., "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2758–2766.
[58] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.
[59] H. Xie, H. Yao, X. Sun, S. Zhou, and S. Zhang, "Pix2Vox: Context-aware 3D reconstruction from single and multi-view images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 2690–2698.
[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.
[61] A. X. Chang et al., "ShapeNet: An information-rich 3D model repository," 2015, *arXiv:1512.03012*.
[62] X. Sun et al., "Pix3D: Dataset and methods for single-image 3D shape modeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2974–2983.
[63] M. Tatarchenko et al., "What do single-view 3D reconstruction networks learn," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3405–3414.
[64] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3D surface construction algorithm," in *Proc. 14th Annu. Conf. Comput. Graph. Interactive Techn.*, 1987, pp. 163–169.
[65] 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.: Annu. Conf. Neural Inf. Process. Syst.*, 2016, pp. 82–90.
[66] J. Wu et al., "MarrNet: 3D shape reconstruction via 2.5D sketches," in *Proc. Adv. Neural Inf. Process. Syst.: Annu. Conf. Neural Inf. Process. Syst.*, 2017, pp. 540–550.
[67] 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.*, 2017, pp. 209–217.
[68] J. Wu et al., "Learning shape priors for single-view 3D completion and reconstruction," in *Proc. 15th Eur. Conf. Comput. Vis.*, 2018, pp. 673–691.
[69] P. O. Pinheiro, N. Rostamzadeh, and S. Ahn, "Domain-adaptive single-view 3D reconstruction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 7637–7646.
[70] M. Rünz et al., "FroDO: From detections to 3D objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14708–14717.
[71] S. R. Richter and S. Roth, "Matryoshka networks: Predicting 3D geometry via nested shape layers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1936–1944.
[72] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[73] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.
[74] L. van der Maaten and G. E. Hinton, "Visualizing non-metric similarities in multiple maps," *Mach. Learn.*, vol. 87, no. 1, pp. 33–55, 2012.
[75] L. van der Maaten, "Accelerating t-SNE using tree-based algorithms," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3221–3245, 2014.