VIPNet: Combining Viewpoint Information and Shape Priors for Instant Multi-View 3D Reconstruction

Weining Ye[1], Zhixuan Li[2], and Tingting Jiang[3]

[1] School of Computer Science, Peking University, Beijing 100871, China
[2] College of Computing and Data Science, Nanyang Technological University, Singapore
[3] National Engineering Research Center of Visual Technology, National Key Laboratory for Multimedia Information Processing, School of Computer Science, National Biomedical Imaging Center, Peking University, Beijing 100871, China

Abstract. While the multi-view 3D reconstruction task has made significant progress, existing methods simply fuse multi-view image features without effectively leveraging available auxiliary information, especially the viewpoint information for guiding and associating features of different views. To this end, we propose to enhance multi-view 3D reconstruction with the power of viewpoint information. Specifically, a simple-yet-effective viewpoint estimator is designed to learn and provide comprehensive viewpoint knowledge for locating and associating learned features from different views. Moreover, to improve the 3D reconstruction quality when 2D images of only very few viewpoints are available, we propose to learn the shape prior knowledge to provide sufficient shape information for compensating the limited 2D observations. Overall, we present **VIPNet**, benefiting from **V**iewpoint **I**nformation and Shape **P**rior learning for high-quality multi-view 3D reconstruction. Extensive experiments validate the effectiveness of the proposed VIPNet, which achieves state-of-the-art performance on challenging datasets and shows well generalization ability in real-world scenarios.

Keywords: Multi-view 3D reconstruction · Viewpoint · Shape Prior

1 Introduction

Multi-view 3D reconstruction task [25,30,34,38,42] aims to recover the geometric structure of 3D scenes from 2D images of different viewpoints. It has emerged as a key component in various applications including augmented reality [39], virtual reality [2], drone navigation [13], and medical image processing [16].

Traditional methods for multi-view 3D reconstruction such as SfM [15] and SLAM [5] are handcrafted to match image features across views, handling well on simple scenarios but perform badly at complex scenes like self-occlusions or irregular shapes. To deal with these limitations, deep learning based approaches

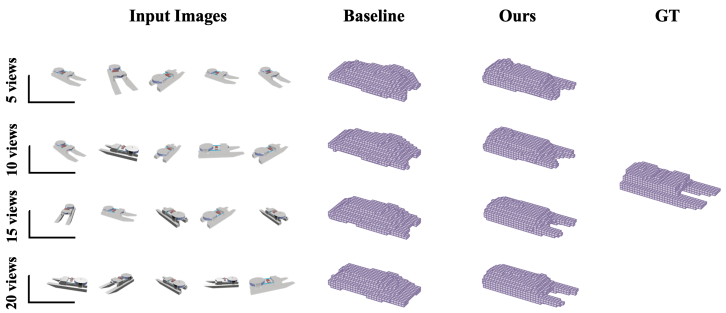

Fig. 1: Visualization of results compared to baseline [42]. Our model which utilizes viewpoint information is able to distinguish between the bow and stern of the watercraft and therefore is able to reconstruct the watercraft more accurately. The more viewpoints there are, the more effective it is.

are designed to learn intrinsic features to describe geometry structures and recover 3D shapes. These approaches can be categorized based on their used 3D representations, including meshes [27,30], point clouds [12,19] and voxels [3,34]. In this work, we focus on reconstructing single 3D object with voxel representation from multiple images.

Existing mainstream deep learning based approaches can be categorized as convolutional neural network (CNN) based [3, 26, 34, 37, 42] and transformer-based [18, 25, 36, 38, 41] approaches. Although these methods achieve fine performance, they still suffer from associating features from different views under complicated scenarios and constructing 3D shapes when only very few viewpoints are available.

Associating features from multiple views poses challenges, especially when the appearance of the 3D object is changed across different views. The lack of explicit feature-to-viewpoint correspondence hinders efficient shape comprehension from the feature representations and accurate 3D reconstruction. Existing methods have not utilized readily available information like viewpoint information or depth data, leading to an unsatisfied performance for multi-view feature fusion. To deal with this issue, we propose to learn the feature-to-viewpoint correspondence from the readily accessible viewpoint information, which provides key guidance to the multi-view feature fusion and helps to distinguish between similar features from different views, thus reconstructing accurate 3D shapes. For example, Fig. 1 shows that baseline method [42] could not distinguish the bow and stern features and reconstruct the similar shape for the two parts, while our method powered by the viewpoint information could understand feature-to-viewpoint correspondence correctly and reconstruct high-quality 3D shapes.

Apart from the multi-view feature association problem, another one is how to deal with performance descending when only 2D images of a few viewpoints are available. Inspired by a single-view 3D reconstruction method [40], we propose to utilize 3D shape prior knowledge for achieving good performance with limited

2D observations. Shape prior is a natural solution as the human reconstructs 3D shapes by retrieving the closely matched 3D model from memory taking 2D images as clues. With the help of the shape prior learning, our method can effectively reconstruct accurate 3D shapes based on 2D observations from very few viewpoints.

Based on the above considerations, we propose **VIPNet**, enhanced with **V**iewpoint **I**nformation and Shape **P**rior learning for accurate multi-view 3D reconstruction to handle complex scenarios and limited viewpoints. Specifically, we propose a simple-yet-effective viewpoint estimator to learn the intrinsic feature of the viewpoint information, which is compact and can guide the model to distinguish confusing 2D features for accurate feature-to-viewpoint associations and high-quality multi-view 3D reconstruction. Moreover, we design a shape-prior-based approach that learns and stores the shape prior knowledge in a dictionary, and retrieves the shape prior from it with 3D rough volume generated by the baseline model as query clues. This approach can deal with the limited consistency between 2D and 3D, thus reconstructing fine 3D shapes with the help of pre-stored shape prior knowledge. Besides, it can also compensate for the viewpoint information which is less significant when the number of viewpoints is limited. Additionally, both the viewpoint feature and the shape prior knowledge feature are combined to enhance the robustness and generalization ability of our basic encoder-decoder structure to achieve better multi-view 3D reconstruction performance under complicated scenarios and when only limited viewpoints are available.

To verify the effectiveness of the above considerations, we implement our method on a CNN-based baseline. The experiments are conducted on the popular synthetic dataset ShapeNet [3, 32] as well as the more challenging real-world dataset Pix3D [21]. Our method not only achieves state-of-the-art results but also demonstrates the generalization ability in real-world scenarios, showing its practical applicability.

Our contributions can be summarized in three major aspects. *First*, we present the VIPNet, which can effectively associate features of different views under complicated scenarios by learning the intrinsic viewpoint knowledge. To the best of our knowledge, this is the first work to enhance the multi-view 3D reconstruction for the voxel representation by using the viewpoint information. *Second*, for the multi-view voxel reconstruction task, we propose an elaborate shape prior learning and retrieving approach to deal with the limited 2D observations when only a few viewpoints are available. *Third*, extensive experiments on ShapeNet and Pix3D demonstrate that VIPNet outperforms the state-of-the-art methods consistently and can generalize well in real-world scenarios.

2 Related Works

2.1 Single-View 3D Reconstruction

Single-view 3D reconstruction is a difficult task due to its limited input views. There are many different representations for 3D models, such as point clouds,

meshes, and voxels. PSGN [4] and 3D-LMNet [14] generate point clouds from single-view images. Pix2Mesh [27] represents the 3D object by a triangular mesh. For voxel representation, there are more methods. 3D-VAE-GAN [31] uses both GAN [6] and VAE [11] to generate 3D models, but it needs class labels to reconstruction. OGN [22] and O-CNN [28] adopt octree to represent higher resolution voxels. To supplement the missing information in the image, Mem3D [40] constructs shape priors which are helpful in recovering the 3D shape of an object, especially when the object is heavily occluded or in a complex environment.

2.2 Multi-View 3D Reconstruction

Traditional methods for multi-view 3D reconstruction such as SfM [15] and SLAM [5] have relied on feature matching between images to derive 3D models. However, they struggle when handling substantial variations in viewpoint, which can result in suboptimal matching quality. Furthermore, the need for manual feature engineering in traditional methods introduces vulnerability to noise and mismatch errors.

In contrast, recent advances have harnessed deep learning to address these challenges. Pixel2Mesh++ [30] predicts a 3D mesh from a coarse mesh by a GCN network. [12,19,29] predict point clouds from multi-view images representing the surface of 3D objects. In this work, we focus on the voxel representation. There are two types of methods based on the voxel representation, CNN-Based and Transformer-Based.

CNN-Based. Methods such as 3D-R2N2 [3] and LSM [9] employ RNN models to fuse image features from multiple viewpoints. However, RNNs are time-consuming and permutation-variant. 3DensiNet [26] utilizes max-pooling to aggregate multi-view image features. This approach often oversimplifies the representation. In contrast, models such as AttSet [37] and Pix2Vox [34] employ attention mechanisms to combine multi-view features, offering a more nuanced understanding of the scene. GARNet [42] further refines Pix2Vox [34] by integrating global image features to aid in 3D voxel generation.

Transformer-Based. EvolT [25] and LegoFormer [36] incorporate Transformer structures within their encoders to amalgamate features from diverse viewpoints, enhancing 3D scene understanding. 3D-RETR [18] utilizes Vision Transformer (ViT) for feature extraction from each view. UMIFormer [41] and LRGT [38] advance this approach, enhancing both intra-view and inter-view feature representations through the power of Transformers.

3 Methodology

The input of our multi-view 3D reconstruction task consists of n images from different viewpoints, denoted as $I_1, I_2, ..., I_n \in R^{H \times W \times 3}$. The output is the final volume $V_f \in [0,1]^{D \times D \times D}$.

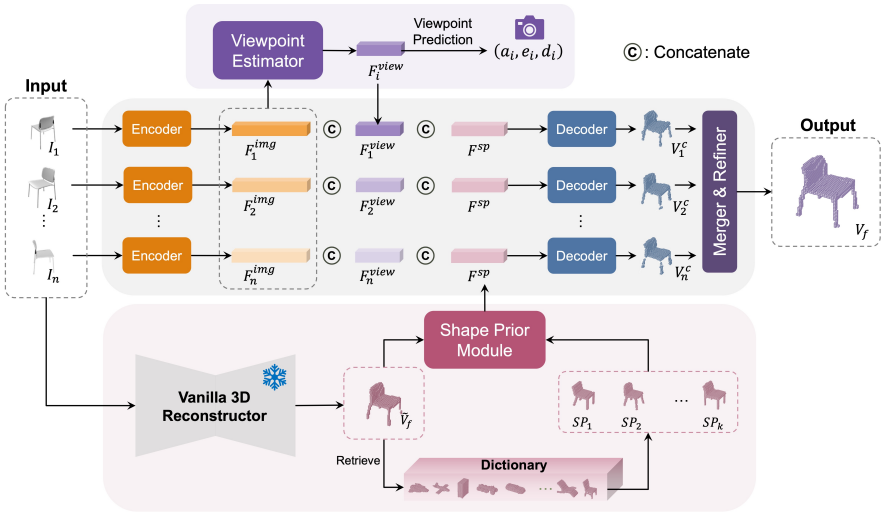

Fig. 2: The architecture of our VIPNet. The main pipeline contains an encoder, a decoder, a merger, and a refiner. The viewpoint feature F_i^{view} is generated by the viewpoint estimator which directly takes F_i^{img} as input. Furthermore, we predict the viewpoint parameters, specifically azimuth, elevation, and distance, denoted as (a_i, e_i, d_i). The rough volume $\tilde{V}_f$ is the final output of the vanilla 3D reconstructor. We use $\tilde{V}_f$ to retrieve k shape priors from the Dictionary. Then the shape prior module takes $\tilde{V}_f$ and shape priors as input and output the shape prior feature F^{sp}. Finally, for each view, we concatenate $F_i^{img}, F_i^{view}, F^{sp}$ together to generate coarse volume V_i^c by Decoder and get final volume V_f after Merger and Refiner.

For each object, the Encoder is utilized to learn and extract essential features F_i^{img} from the input image. Next, the viewpoint estimator learns the intrinsic viewpoint knowledge which is embedded as F_i^{view} to provide viewpoint information for distinguishing the learned image feature F_i^{img} from similar counterparts. Besides, the Vanilla 3D Reconstructor with frozen parameters is utilized for obtaining the rough volume $\tilde{V}_f$ of each object, which is further used to retrieve shape priors from the Dictionary and create the shape prior feature F^{sp}. Then the Decoder reconstructs the 3D shape based on the comprehensive learning from F_i^{img}, F_i^{view}, and F^{sp} to achieve accurate 3D reconstruction V_i^c of this viewpoint. Finally, a Merger and Refiner are proposed to integrate the complementary reconstructed 3D shapes of all viewpoints to obtain high-quality 3D reconstruction V_f.

3.1 Main Architecture

The main architecture of our proposed method consists of an Encoder, Decoder, Merger, and Refiner. Their details are introduced in the following.

Encoder. The encoder is designed to extract the essential features of the input image, which includes rich information for deducting the shape and gesture of the 3D object. The encoder is composed of the first three convolutional blocks from ResNet [7], along with three additional convolutional layers, resulting in feature maps F_i^{img} of size $256 \times 7 \times 7$ for each input image I_i.

Decoder. The decoder is proposed to reconstruct the 3D shape comprehensively based on the learned knowledge from the image feature F_i^{img}, viewpoint feature F_i^{view}, and the shape prior feature F^{sp}. The decoder comprises two parts, D_1 and D_2, including one and four transposed convolutional layers, respectively. Initially, image features are passed through D_1 to produce $(F_1^{'}, F_2^{'}, ..., F_n^{'})$. These features are then fused into a global feature $F_G^{'}$ using spatial attention and channel attention. Subsequently, all these n+1 features are processed through D_2 to yield the voxel predictions for each viewpoint $(\tilde{V}_1^c, \tilde{V}_2^c, ..., \tilde{V}_n^c)$ as well as the final features $(F_1^{''}, F_2^{''}, ..., F_n^{''}, F_G^{''})$.

Merger. Since different viewpoints contribute to the final 3D reconstructed shape unequally, the merger is designed to predict a score map for each viewpoint to indicate the importance of this viewpoint, which is further utilized as guidance to fuse the reconstructed 3D shapes V_i^c as a completed yet coarse volume $\tilde{V}_M^c$. Specifically, the input of the merger is the concatenation of two feature maps $(F_i^{''}, F_G^{''} - F_i^{''})$. Then the coarse volume $\tilde{V}_M^c$ is obtained after predicting the score maps with a softmax layer and fusing the $\tilde{V}_i^c$ by the score maps.

Refiner. The refiner is proposed to improve the quality of the fused coarse 3D volume $\tilde{V}_M^c$ to obtain a fine and accurate final prediction. Specifically, the refiner is a UNet [17] network structure combining with ResNet [7] structure, further refining $\tilde{V}_M^c$ to generate the final prediction $\tilde{V}_f$.

3.2 Viewpoint Information

To enhance the semantic information related to the viewpoint within the image feature, we directly use the image feature as input to extract viewpoint features as shown in Fig. 3. Then, a single fully connected network is employed to predict viewpoint parameters, including azimuth, elevation, and distance.

Viewpoint Feature. In order to extract viewpoint features from the image feature F_i^{img}, we start by applying 2D Average Pooling to each channel of the image feature, resulting in a 256-dimensional vector. Subsequently, a two-layer fully connected network, with an output dimension of 1024 for each layer, is utilized to obtain a final 1024-dimensional viewpoint feature F_i^{view}, as shown in Fig. 3. This feature is extracted from the image feature and contains semantic information related to the viewpoint. When concatenated with the image feature, it effectively aids the Decoder and Merger in better reconstruction and fusion based on the viewpoint information.

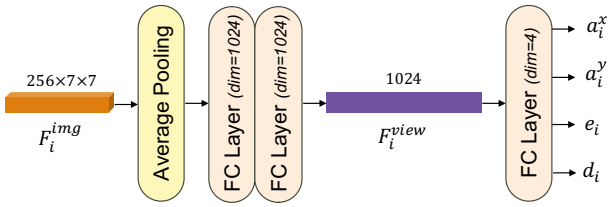

Fig. 3: Architecture of the Viewpoint Estimator

Viewpoint Prediction. To ensure that the viewpoint feature genuinely contains the necessary viewpoint information, we further process the obtained viewpoint feature through a single fully connected network to predict viewpoint parameters, specifically azimuth, elevation, and distance, denoted as (a_i, e_i, d_i), respectively. Azimuth ranges from 0 to 360, elevation from 0 to 30, and distance from 0 to 1. To address the discontinuity issue in azimuth prediction within the domain [0, 360], we convert it into Cartesian coordinates (a_i^x, a_i^y), where $a_i = atan2(a_i^x, a_i^y)$. This transformation results in a final network prediction of four values: (a_i^x, a_i^y, e_i, d_i). The viewpoints predicted are directly supervised by the ground truth viewpoints which are already provided by [3].

Note that we only use ground truth viewpoints annotations to supervise our viewpoint prediction module for training. For testing, no viewpoints are given.

3.3 Shape Prior Module

To further enhance the model's performance when dealing with a limited number of viewpoints, we propose to utilize shape prior with a designed dictionary for storing and providing sufficient shape knowledge.

For multi-view input images, we first use the pre-trained vanilla 3D reconstructor with frozen parameters to produce a rough 3D model $\tilde{V}_f$. This 3D volume is decoded directly from the features of the images without shape priors. Within the Dictionary, we retrieve several similar volumes to serve as shape priors. Finally, we use Multi-Head Attention and one fully connected (FC) layer to merge these volumes into the ultimate shape prior feature. With this proposed shape prior approach, our method can effectively leverage pre-stored shape prior knowledge to deal with complicated scenarios especially when only very few viewpoints are provided and lack sufficient 2D observations.

Dictionary. We construct the Dictionary based on the Ground Truth (GT) volumes of the training dataset. Specifically, the Dictionary consists of a total of m volumes, initially empty. For each sample in the training set, if its Intersection over Union (IoU) with all the volumes currently in the Dictionary is less than a threshold δ, then we add it to the Dictionary. If the Dictionary is not full, we add the volume directly. Otherwise, we remove the volume that entered the Dictionary earliest and then add the current volume. This process ensures that the Dictionary contains a representative set of volumes within its capacity limit.

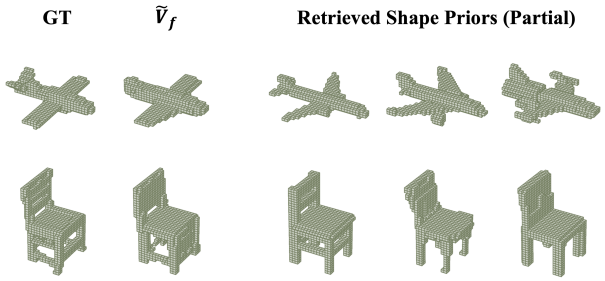

Fig. 4: Visualization of retrieved shape priors

The steps described above are performed before training, and we will save the Dictionary locally for future use. During subsequent training and testing, we directly load the saved Dictionary and retrieve shape priors. Specifically, we first calculate the Intersection over Union (IoU) between the rough volume $\tilde{V}_f$ and all the volumes in the Dictionary, and then we select the top k entries with the highest IoU as the final set of shape priors. Fig. 4 shows two examples, we exhibit partial shape priors retrieved from the Dictionary.

Multi-Head Attention. With the Dictionary in place, we use the rough estimation $\tilde{V}_f$ to find the k shape priors in the Dictionary, denoted as $SP_1, SP_2, ..., SP_k$. Next, we need to merge them into a single shape prior feature. To accomplish this, we apply the Multi-Head Attention mechanism.

As shown in Fig. 5, we treat $\tilde{V}_f$ as the Query and $(SP_1, SP_2, ..., SP_k)$ as the Key and Value. Initially, we employ a volume encoder which consists of four layers of three-dimensional convolutional networks to separately extract voxel features for $\tilde{V}_f$, $SP_1, SP_2, ..., SP_k$, resulting in q, k_i, and v_i, $i = 1, 2, ..., k$. Then we use three separate linear layers parameterized by W_q, W_k, W_v to get query, key, and value embedding Q, K, V. Subsequently, the Q, K, V are fed into the Multi-Head Attention (MHA) [24] module and the Layer Normalization [1] module to perform cross-attention. We apply residual connection to fuse the input query embedding with the output to get enhanced feature F_Q. The final shape prior feature F^{sp} is obtained after a fully connected (FC) layer and Layer Normalization with residual connection. The whole pipeline is formulated as:

$$q = \text{VolEnc}(\tilde{V}_f), \quad k_i, v_i = \text{VolEnc}(SP_i) \tag{1}$$

$$Q = qW_q, K = k_i W_k, V = v_i W_v \tag{2}$$

$$F_Q = Q + \text{LN}(\text{MHA}(Q, K, V)) \tag{3}$$

$$F^{sp} = F_Q + \text{LN}(\text{FC}(F_Q)) \tag{4}$$

3.4 Loss Function

For the viewpoint prediction (a_i, e_i, d_i) and their corresponding Ground Truth values $(\hat{a}_i, \hat{e}_i, \hat{d}_i)$, we first convert azimuth a_i and elevation e_i into coordinates

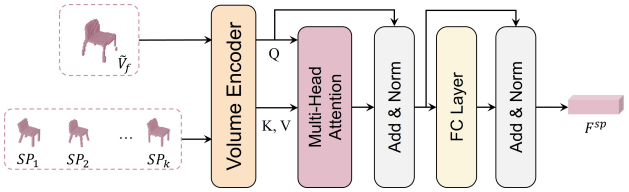

Fig. 5: Architecture of the Shape Prior Module

on the unit circle while leaving distance unchanged. Then, we calculate the final viewpoint loss function using the L_2 loss for all these parameters and take the average of all views.

$$\mathcal{D}(u, v) = \|\cos u - \cos v\|^2 + \|\sin u - \sin v\|^2 \tag{5}$$

$$\mathcal{L}_{view} = \frac{1}{n}\sum_{i=1}^{n}\left[\mathcal{D}(a_i, \hat{a}_i) + \mathcal{D}(e_i, \hat{e}_i) + \|d_i - \hat{d}_i\|^2\right] \tag{6}$$

Following the previous work, we also apply binary cross-entropy(BCE) loss to supervise the coarse volume V_M^c and the final volume V_f. The BCE loss between predicted volume V and ground truth volume V^g is formulated as:

$$\mathcal{L}_{vox}(V) = \frac{1}{N}\sum_{i=1}^{N}[V_i^g \log V_i + (1 - V_i^g)\log(1 - V_i)] \tag{7}$$

where N denotes the number of voxels in volume. V_i and V_i^g represent the predicted occupancy and the corresponding ground truth.

Therefore, the complete loss function of our model is defined as:

$$\mathcal{L} = \mathcal{L}_{vox}(V_M^c) + \mathcal{L}_{vox}(V_f) + 0.1 \cdot \mathcal{L}_{view}. \tag{8}$$

4 Experiments

4.1 Datasets and Metrics

We assess the performance of all the models using the ShapeNet [32] dataset. Following [3], our evaluation focuses on a specific subset of ShapeNet, which comprises 13 major categories and encompasses a total of 43,783 3D models. Besides, following [34, 35], we evaluate our model on challenging real-world dataset Pix3D [21], which contains 2894 untruncated and unoccluded single-view chair images. Following the current advanced works, we take the mean Intersection-over-Union (IoU) and the F-Score@1% [23, 35] as our metrics.

4.2 Implementation Details

The input image resolution and output voxel resolution are set as $H = W = 224, D = 32$. We adopt an Adam [10] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to

Table 1: Comparison of multi-view 3D reconstruction on ShapeNet using IoU / F-Score@1%. Following GARNet [42], our VIPNet and VIPNet+ take 3 and 8 as the maximum number of input views during training, respectively. Best in **bold**, second underlined.

Method	1 view	2 views	3 views	5 views	8 views	12 views	20 views
3D-R2N2 [3]	0.560 / 0.351	0.603 / 0.368	0.617 / 0.372	0.634 / 0.382	0.635 / 0.383	0.636 / 0.382	0.636 / 0.383
AttSets [37]	0.642 / 0.395	0.662 / 0.418	0.670 / 0.426	0.677 / 0.432	0.685 / 0.444	0.688 / 0.445	0.693 / 0.448
Pix2Vox++ [35]	0.670 / **0.436**	0.695 / 0.452	0.704 / 0.455	0.711 / 0.458	0.715 / 0.459	0.717 / 0.460	0.719 / 0.462
GARNet [42]	0.673 / 0.418	0.705 / 0.455	0.716 / 0.468	0.726 / 0.479	0.731 / 0.486	0.734 / 0.489	0.737 / 0.492
GARNet+	0.655 / 0.399	0.696 / 0.446	0.712 / 0.465	0.725 / 0.481	0.733 / 0.491	0.737 / 0.498	0.742 / 0.504
VIPNet	**0.676** / 0.421	**0.714** / **0.464**	0.727 / 0.479	0.737 / 0.491	0.742 / 0.498	0.745 / 0.502	0.747 / 0.505
VIPNet+	0.669 / 0.409	0.713 / 0.460	**0.729** / **0.480**	**0.742** / **0.496**	**0.750** / **0.506**	**0.754** / **0.512**	**0.757** / **0.516**

Table 2: Effect of Viewpoint Information and Shape Priors. VP, VF, and SP represent predicting viewpoints from image features, concatenating viewpoint features F_i^{view} and image features, and concatenating shape prior features and image features, respectively. The results are IoU performance testing on the ShapeNet dataset. All the models take 8 as the maximum number of input views during training.

VP	VF	SP	1 view	2 views	5 views	8 views	12 views	20 views
			0.655	0.696	0.725	0.733	0.737	0.742
✓			0.659	0.703	0.734	0.743	0.748	0.753
✓	✓		0.661	0.705	0.735	0.744	0.749	0.753
		✓	0.665	0.707	0.735	0.743	0.747	0.751
✓		✓	0.666	0.709	0.738	0.747	0.751	0.755
✓	✓	✓	**0.669**	**0.713**	**0.742**	**0.750**	**0.754**	**0.757**

train the multi-view 3D reconstruction with a batch size of 32 for 200 epochs. The learning rate is 2e-4 initially and reduces to half after [40, 60, 80, 100, 140, 180] epochs sequentially. For all the experiments, we take the pre-trained GARNet+ [42] as the Vanilla 3D Reconstructor. The capacity of the Dictionary is 4000 and the number of retrieved shape priors k is set to 8. We use the same training strategy as [42], named dynamic two-stage training strategy. Eventually, following [38, 41, 42], we provide two models respectively setting the maximum number of input views to 3 and 8 during training, named VIPNet and VIPNet+, which are all trained on 2 Tesla V100 for about 2 days and 5 days respectively. The fixed threshold for binarizing the probabilities is set as 0.3.

4.3 Evaluation on the ShapeNet Dataset

We compare our proposed multi-view 3D reconstruction methods VIPNet and VIPNet+ with the current SOTA methods on the ShapeNet dataset. As shown in Tab. 1, compared to other CNN-based methods, our VIPNet and VIPNet+ dominate in almost all metrics. We also compare the visual results shown in Fig. 6. For chair reconstruction, our VIPNet+ has better reconstruction for both left and right arms due to our shape prior module and viewpoint information. For the display reconstruction, our VIPNet+ can reconstruct a more accurate and complete screen.

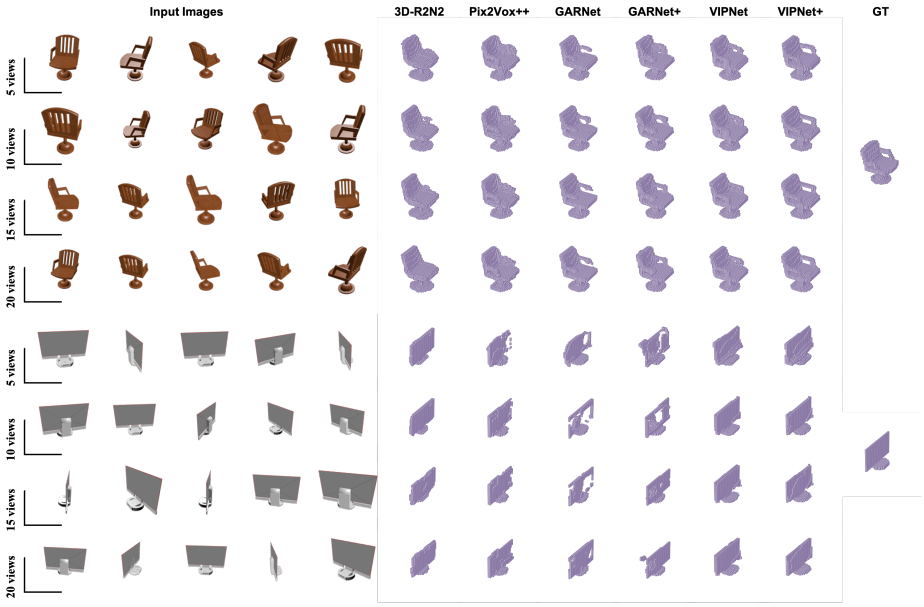

Fig. 6: Visualization of Multi-View 3D reconstruction with other methods on ShapeNet when facing 5, 10, 15, 20 as input views number.

Table 3: The quality of retrieved shape prior when testing on ShapeNet.

IoU(SP,GT)	≥ 0.2	≥ 0.3	≥ 0.4	≥ 0.5
1-view	98.1%	94.1%	86.6%	74.7%
3-view	98.7%	95.4%	88.6%	77.0%
20-view	98.8%	95.7%	89.1%	77.6%

To evaluate the quality of the retrieved shape prior, we calculate the IoU between the shape prior SP_1 and the ground truth volume for each test sample in ShapeNet. We count the proportion of test samples with IoU greater than 0.2, 0.3, 0.4, and 0.5, respectively. Tab. 3 shows the results. Only with single-view input, about 75% of the test samples can retrieve the shape prior from the Dictionary with IoU greater than 0.5. With 20 views input, about 99% of the test samples can retrieve the shape prior with IoU greater than 0.2.

To verify the accuracy of the viewpoint prediction, we calculated the viewpoint accuracy on the test set of ShapeNet. Specifically, we first convert the 'azimuth' and 'elevation' into Cartesian coordinates. When the distance between the predicted point and the ground truth point is less than 0.1, and the difference in 'distance' is also less than 0.1, we consider the viewpoint prediction to be correct. Ultimately, we achieved a 98% accuracy rate on the test set of ShapeNet.

Table 4: IoU and F-score@1% results of single-view reconstruction on Pix3D.

	Pix2Vox++	GARNet	VIPNet
IoU / F-score@1%	0.279 / 0.113	0.291 / 0.116	**0.300** / **0.140**

Table 5: IoU results of different capacities of the Dictionary testing on ShapeNet. All the models take 3 as the maximum number of input views during training. m represents the capacity of the dictionary.

	1 view	2 views	5 views	20 views
$m = 1000$	0.668	0.707	0.732	0.746
$m = 2000$	0.676	0.713	0.734	0.745
$m = 4000$	**0.676**	**0.714**	**0.737**	**0.747**

This demonstrates that our viewpoint prediction is effective and indicates that our image features indeed contain viewpoint information.

4.4 Evaluation on the Pix3D Dataset

To evaluate our proposed methods on real-world images, we conduct single-view reconstruction on the Pix3D dataset. Following the previous methods [34, 35], we generate a new training dataset of ShapeNet. Specifically, we use the pipeline of RenderForCNN [20] to render 60 images for each 3D model of the chair in the ShapeNet dataset and then synthesize images with random background from the SUN dataset [33]. Tab. 4 shows the performance compared to other methods. With the help of viewpoint information and shape priors, VIPNet still achieves satisfactory results in challenging real-world scenarios. The IoU result of our method is better than all the other methods. For F-score@1%, our method also achieves SOTA performance, demonstrating robustness in real-world scenarios. Moreover, Fig. 7 shows the visualization results on Pix3D. Our model can generate more accurate 3D shapes compared to other methods.

4.5 Ablation Study

Ablation studies are conducted regarding our method only retaining the Encoder, Decoder, Merger, and Refiner as the baseline model. We evaluate the effect of the viewpoint information and shape priors on the performance of multi-view 3D reconstruction. All the experiments are conducted on the ShapeNet dataset.

Viewpoint Information. To evaluate the effect of viewpoint information, we design three models: not using any viewpoint information, only using viewpoint prediction(VP), and using viewpoint prediction (VP) as well as concatenating the viewpoint features(VF) F_i^{view} with image features F_i^{img}. We can see from

Table 6: IoU results on ShapeNet when building Dictionary on three random orders. All the models take 3 as the maximum number of input views during training. The results are shown in "mean ± std" format.

1 view	2 views	5 views	20 views
0.6771 ± 0.0009	0.7147 ± 0.0009	0.7371 ± 0.0004	0.7479 ± 0.0005

Table 7: IoU results of different numbers of retrieved shape priors testing on ShapeNet. All the models take 8 as the maximum number of input views during training. k represents the number of retrieved shape priors.

	1 view	2 views	5 views	20 views
$k = 1$	0.666	0.710	0.738	0.754
$k = 8$	**0.669**	**0.713**	**0.742**	**0.757**
$k = 16$	0.668	0.711	0.740	0.755

Tab. 2 that only applying viewpoint prediction without concatenating viewpoint features also improves the performance, indicating that viewpoint prediction indeed enhances the semantic representation of image features. Besides, concatenating the viewpoint feature (VF) further improves the performance. Fig. 1 also indicates the effectiveness of viewpoint information.

Shape Prior Module. We evaluate the importance of the Shape Prior module. Tab. 2 shows the shape prior module indeed improves the multi-view 3D reconstruction performance whether with viewpoint information or not.

We also evaluate whether the capacity and building order of the Dictionary affect the performance of reconstruction. Tab. 5 indicates that the IoU results are higher with the larger capacity of the Dictionary, and the performance is almost converged when $m = 4000$. To evaluate the impact of the order of traversing the dataset to build a Dictionary, we tried to generate three dictionaries by different random traversing orders and train the network with these Dictionaries separately. Tab. 6 shows the test results. We can see that it does not affect the final performance. Tab. 7 shows the number of retrieved shape priors does not affect the performance much and the results indicate that 8 is the best.

Furthermore, we explore different ways to fuse the shape priors retrieved from the Dictionary. For instance, we can directly use the Top-1 retrieved shape as the shape prior. We can also use average fusion or the LSTM [8] network to encoder retrieve shapes into shape prior features. Besides, we also try directly taking the rough volume $\tilde{V}_f$ as the shape prior. Tab. 8 shows the results. The model with MHA performs best on each number of input views.

Effect on Different Numbers of Input Views. We evaluate the effect of viewpoint information and shape prior when testing on different numbers of

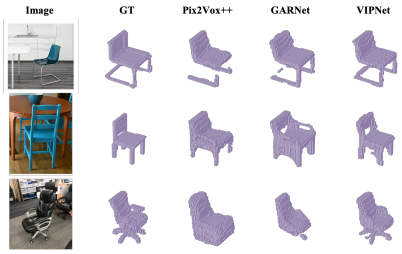

Fig. 7: Visualization of reconstruction on Pix3D datasets. The input is a single image with complex background.

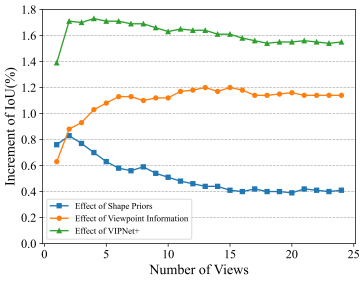

Fig. 8: The improvement of viewpoint information and shape priors when testing on each number of input views.

Table 8: IoU results of different shape prior fusion modules testing on ShapeNet. All the models take 3 as the maximum number of input views during training.

	1 view	2 views	5 views	20 views
$\tilde{V}_f$	0.670	0.710	0.734	0.746
1 SP	0.673	0.709	0.729	0.739
Average	0.674	0.709	0.730	0.741
LSTM	0.675	0.710	0.732	0.744
MHA	**0.676**	**0.714**	**0.737**	**0.747**

input views. Fig. 8 shows the results. Notably, the improvement in viewpoint information is more pronounced when dealing with a higher number of input viewpoints, whereas the enhancement from shape priors is more prominent when dealing with fewer input viewpoints.

5 Conclusion

In conclusion, we propose a novel multi-view 3D reconstruction network, VIPNet. It combines viewpoint information and shape priors to improve the performance on whether few or high numbers of views. Our experiments have shown that viewpoint information plays a critical role in 3D reconstruction, especially in a surplus of input viewpoints. While, the shape prior module further enhances the network's ability to reconstruct accurate shapes, which is particularly valuable with limited input viewpoints. Assisted by viewpoint information and shape priors, our method exhibits robust generalization ability in real-world scenarios.

Acknowledgments. This work was partially supported by Sino-German Center (M 0187) and the Natural Science Foundation of China under contract 62088102. We also acknowledge High-Performance Computing Platform of Peking University for providing computational resources.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer Normalization. arXiv preprint arXiv:1607.06450 (2016)
2. Bruno, F., Bruno, S., De Sensi, G., Luchi, M.L., Mancuso, S., Muzzupappa, M.: From 3D Reconstruction to Virtual Reality: A Complete Methodology for Digital Archaeological Exhibition. Journal of Cultural Heritage **11**(1), 42–49 (2010)
3. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3D-R2N2: A Unified Approach for Single and Multi-View 3D Object Reconstruction. In: Proceedings of the European Conference on Computer Vision. pp. 628–644. Springer (2016)
4. Fan, H., Su, H., Guibas, L.J.: A Point Set Generation Network for 3D Object Reconstruction from a Single Image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 605–613 (2017)
5. Fuentes-Pacheco, J., Ruiz-Ascencio, J., Rendón-Mancha, J.M.: Visual Simultaneous Localization and Mapping: A Survey. Artificial Intelligence Review **43**, 55–81 (2015)
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. Advances in Neural Information Processing Systems **27** (2014)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
8. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation **9**(8), 1735–1780 (1997)
9. Kar, A., Häne, C., Malik, J.: Learning a Multi-View Stereo Machine. Advances in neural information processing systems **30** (2017)
10. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. ArXiv Preprint ArXiv:1312.6114 (2013)
12. Lin, C.H., Kong, C., Lucey, S.: Learning Efficient Point Cloud Generation for Dense 3D Object Reconstruction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
13. Lu, Y., Xue, Z., Xia, G.S., Zhang, L.: A Survey on Vision-Based UAV Navigation. Geo-spatial Information Science **21**(1), 21–32 (2018)
14. Mandikal, P., Navaneet, K., Agarwal, M., Babu, R.V.: 3D-LMNet: Latent Embedding Matching for Accurate and Diverse 3D Point Cloud Reconstruction from a Single Image. ArXiv Preprint ArXiv:1807.07796 (2018)
15. Özyeşil, O., Voroninski, V., Basri, R., Singer, A.: A Survey of Structure from Motion*. Acta Numerica **26**, 305–364 (2017)
16. Razzak, M.I., Naz, S., Zaib, A.: Deep Learning for Medical Image Processing: Overview, Challenges and the Future. Classification in BioApps: Automation of Decision Making pp. 323–350 (2018)
17. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
18. Shi, Z., Meng, Z., Xing, Y., Ma, Y., Wattenhofer, R.: 3D-RETR: End-to-End Single and Multi-View 3D Reconstruction with Transformers. In: 32nd British Machine Vision Conference (BMVC 2021). p. 405 (2021)

19. Sridhar, S., Rempe, D., Valentin, J., Sofien, B., Guibas, L.J.: Multiview Aggregation for Learning Category-Specific Shape Reconstruction. Advances in Neural Information Processing Systems **32** (2019)
20. Su, H., Qi, C.R., Li, Y., Guibas, L.J.: Render for CNN: Viewpoint Estimation in Images Using CNNs Trained with Rendered 3D Model Views. In: Proceedings of the IEEE International Conference on Computer Vision (December 2015)
21. Sun, X., Wu, J., Zhang, X., Zhang, Z., Zhang, C., Xue, T., Tenenbaum, J.B., Freeman, W.T.: Pix3D: Dataset and Methods for Single-Image 3D Shape Modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2974–2983 (2018)
22. Tatarchenko, M., Dosovitskiy, A., Brox, T.: Octree Generating Networks: Efficient Convolutional Architectures for High-Resolution 3D Outputs. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2088–2096 (2017)
23. Tatarchenko, M., Richter, S.R., Ranftl, R., Li, Z., Koltun, V., Brox, T.: What do Single-View 3D Reconstruction Networks Learn? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3405–3414 (2019)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is All you Need. Advances in Neural Information Processing Systems **30** (2017)
25. Wang, D., Cui, X., Chen, X., Zou, Z., Shi, T., Salcudean, S., Wang, Z.J., Ward, R.: Multi-View 3D Reconstruction with Transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5722–5731 (2021)
26. Wang, M., Wang, L., Fang, Y.: 3DensiNet: A Robust Neural Network Architecture towards 3D Volumetric Object Prediction from 2D Image. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 961–969 (2017)
27. Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y.G.: Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images. In: Proceedings of the European Conference on Computer Vision. pp. 52–67 (2018)
28. Wang, P.S., Liu, Y., Guo, Y.X., Sun, C.Y., Tong, X.: O-cnn: Octree-based convolutional neural networks for 3d shape analysis. ACM Transactions On Graphics **36**(4), 1–11 (2017)
29. Wei, Y., Liu, S., Zhao, W., Lu, J.: Conditional Single-View Shape Generation for Multi-View Stereo Reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9651–9660 (2019)
30. Wen, C., Zhang, Y., Li, Z., Fu, Y.: Pixel2Mesh++: Multi-View 3D Mesh Generation via Deformation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1042–1051 (2019)
31. Wu, J., Zhang, C., Xue, T., Freeman, B., Tenenbaum, J.: Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling. Advances in Neural Information Processing Systems **29** (2016)
32. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3D ShapeNets: A Deep Representation for Volumetric Shapes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1912–1920 (2015)
33. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: SUN Database: Large-Scale Scene Recognition from Abbey to Zoo. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 3485–3492 (2010). https://doi.org/10.1109/CVPR.2010.5539970
34. Xie, H., Yao, H., Sun, X., Zhou, S., Zhang, S.: Pix2vox: Context-Aware 3D Reconstruction from Single and Multi-View Images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2690–2698 (2019)

35. Xie, H., Yao, H., Zhang, S., Zhou, S., Sun, W.: Pix2Vox++: Multi-Scale Context-Aware 3D Object Reconstruction from Single and Multiple images. International Journal of Computer Vision **128**(12), 2919–2935 (2020)
36. Yagubbayli, F., Wang, Y., Tonioni, A., Tombari, F.: Legoformer: Transformers for Block-by-block Multi-View 3D Reconstruction. arXiv preprint arXiv:2106.12102 (2021)
37. Yang, B., Wang, S., Markham, A., Trigoni, N.: Robust Attentional Aggregation of Deep Feature Sets for Multi-View 3D Reconstruction. International Journal of Computer Vision **128**, 53–73 (2020)
38. Yang, L., Zhu, Z., Lin, X., Nong, J., Liang, Y.: Long-Range Grouping Transformer for Multi-View 3D Reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18257–18267 (2023)
39. Yang, M.D., Chao, C.F., Huang, K.S., Lu, L.Y., Chen, Y.P.: Image-Based 3D Scene Reconstruction and Exploration in Augmented Reality. Automation in Construction **33**, 48–60 (2013)
40. Yang, S., Xu, M., Xie, H., Perry, S., Xia, J.: Single-View 3D Object Reconstruction from Shape Priors in Memory. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3152–3161 (2021)
41. Zhu, Z., Yang, L., Li, N., Jiang, C., Liang, Y.: UMIFormer: Mining the Correlations between Similar Tokens for Multi-View 3D Reconstruction. arXiv preprint arXiv:2302.13987 (2023)
42. Zhu, Z., Yang, L., Lin, X., Yang, L., Liang, Y.: GARNet: Global-aware Multi-View 3D Reconstruction Network and the Cost-Performance Tradeoff. Pattern Recognition **142**, 109674 (2023)