The Visual Computer (2023) 39:6237–6247
https://doi.org/10.1007/s00371-022-02725-6

ORIGINAL ARTICLE

IV-Net: single-view 3D volume reconstruction by fusing features of image and recovered volume

Beibei Sun[1] · **Ping Jiang**[1] · **Dali Kong**[1] · **Ting Shen**[1]

Accepted: 5 November 2022 / Published online: 23 November 2022

Abstract
Single-view 3D reconstruction aims to recover the 3D shape from one image of an object and has attracted increasingly attention in recent years. Mostly, previous works are devoted to learning a mapping from 2 to 3D, and lack of spatial information of objects will cause inaccurate reconstruction on the details of objects. To address this issue, for single-view 3D reconstruction, we propose a novel voxel-based network by fusing features of image and recovered volume, named IV-Net. By a pre-trained baseline, it achieves image feature and a coarse volume from each image input, where the recovered volume contains spatial semantic information. Specially, the multi-scale convolutional block is designed to improve 2D encoder by extracting multi-scale image information. To recover more accurate shape and details of the object, an IV refiner is further used to reconstruct the final volume. We conduct experimental evaluations on both synthetic ShapeNet dataset and real-world Pix3D dataset, and results of comparative experiments indicate that our IV-Net outperforms state-of-the-art approaches about accuracy and parameters.

Keywords Single-view 3D reconstruction · Multi-scale convolution · Deep learning · Residual convolutional neural network

1 Introduction

3D reconstruction generates a 3D shape of an object from single or more 2D images, and it plays an important role in various applications, including medical image processing [1], virtual reality [2], CAD [3], human detection [4], etc. 3D reconstruction has been tackled using conventional computer vision algorithms [5, 6].

But these traditional algorithms require prior knowledge assumptions and sophisticated hardware and are thus not practical in many scenarios. Recently, deep learning has shown powerful advantages in computer vision. So many researchers were prompted to learning-based methods for 3D reconstruction. For multi-view 3D reconstruction, some works pay more attention to matching view features extracted from different views of an object [7–9]. Compared with multi-view reconstruction, single-view 3D reconstruction is more difficult and with some troubles, such as self-occlusion and the absence of sufficient object information from different angles. Therefore, it is necessary to propose a single-view 3D reconstruction algorithm with higher reconstruction accuracy.

Generally, the reconstructed 3D shape of an object can be represented as volume [7–14], mesh [15], or point clouds [16–18], etc. For single-view voxel-based reconstruction, there are several learning-based networks that were proposed to settle the task. For example, some approaches firstly extract view feature from single input image and then transform the view feature into 3D representations [7–9]. 3D VAE-GAN [10] is an adversarial learning-based network. Then, some methods also utilize transformers to perform end-to-end single-image 3D reconstruction [11, 12]. The above methods do not explicitly consider the spatial semantic information of objects, which leads to the inaccuracy or incompleteness of the reconstructed volume. Considering the reconstruction accuracy, parameters and optimization convergence speed of our model, we choose to improve AttSets [8] and propose a better reconstruction framework.

In this paper, for single-view 3D volume reconstruction, we present a novel framework IV-Net. IV-Net reconstructs a refined volume by fusing features of image and volume recovered from baseline and contains two main modules:

✉ Ping Jiang
hfutus@hfut.edu.cn

[1] School of Mathematics, Hefei University of Technology, Hefei 230000, China

a baseline and an IV refiner. For each single-view input, the baseline module is pre-trained to generate a relatively reliable 3D volume, which supplements certain spatial information. Then, based on the pre-trained baseline, an IV refiner further generates a better-reconstructed volume, where the details and shape of an object can be better predicted due to fusing image feature with spatial information.

Our main contributions are as follows:

(1) We construct a unified refiner network for single-view 3D reconstruction, namely IV-Net. It shows the advantages of recovering the shape and details of an object and has universal and adaptable application prospect.
(2) We present a multi-scale convolutional block to extract multi-scale information for enhancing learning ability of 2D encoder.
(3) We construct a residual convolutional neural network as 3D encoder to extract efficiently spatial feature of the recovered volume.
(4) Experimental results on both ShapeNet and Pix3D datasets demonstrate that IV-Net improves the reconstruction quality and performs favorably compared with state-of-the-art methods.

This paper consists of five sections. Section 2 introduces the related work. Section 3 discusses our framework and loss functions. The datasets, evaluation metrics and results of comparative experiments are shown in Sect. 4. Section 5 presents the conclusion and the prospect for future research.

2 Related work

Predicting the 3D shape from a 2D image is a challenging and ill-posed problem. Recently, with the availability of large-scale datasets, there are some learning-based networks presented for single-view reconstruction.

With a limited memory budget, OGN [13] utilizes octree to represent high-resolution 3D reconstructed volumes. Matryoshka networks [14] decompose the 3D shape of an object into nested shape layers and are better than octree-based methods. With the success of generative adversarial networks (GANs) [19] and its variations, 3D VAE-GAN [10] generates a volume from a single-view input by using GAN and variational autoencoders (VAEs). Besides, 3D-R2N2 [7], AttSets [8], and Pix2Vox++ [9], all based on encoder–decoder, firstly encode the single-view input images to fixed-size feature vectors and then pass them into 3D decoder to decode 3D representations. In particular, AttSets [8] chooses the encoder–decoder of 3D-R2N2 [7] and SilNet [20] as its two base nets. Moreover, researchers have also used transformers for 3D volume reconstruction [11, 12]. However, those voxel-based works, without considering spatial information of objects, reconstruct comparatively inaccurate volumes in detail and shape. And different from voxel-based representation, for a given single-view image, Pixel2Mesh [15] applies a graph convolutional network to generate a 3D triangular mesh, and PSGN [16] and 3D-LMNet [17] generate point representations.

3 Methodology

The proposed IV-Net focuses on reconstructing a 3D volume of size 32^3 from a single-view image and contains two-step optimization modules: a baseline and an IV refiner, as illustrated in Fig. 1. Firstly, for each single-view image input, baseline is pre-trained to obtain image feature and a coarse volume, which supplements additional spatial information. Then, IV refiner is further trained to generate a more accuracy volume. Next, we will introduce these two modules in detail, respectively.

3.1 Baseline

Our baseline model includes of three components: a 2D encoder, a latent feature processing (LFP) module and a 3D decoder. From each single-view input, the baseline is pre-trained to get image feature and a coarse 3D volume of size 32^3. Considering the reconstruction accuracy, parameters and optimization convergence speed of our model, we improve the AttSets [8] network as our baseline. Specially, in our paper, its encoder–decoder is based on 3D-R2N2 [7]. The encoder and decoder of 3D-R2N2 are standard residual convolutional neural networks (CNNs), and they can enhance and accelerate the optimization process for very deep networks by adding residual connections between standard convolution layers.

3.1.1 2D Encoder

From each single-view $127 \times 127 \times 3$ image input, 2D encoder gains a fixed 1×1024 image feature vector $\mathbf{z}$, as shown in Fig. 2. Our 2D encoder is based on the 2D encoder of 3D-R2N2 [7], which mainly uses fixed 3×3 convolution in convolution layers. And to extract multi-scale image feature, we design multi-scale convolutional (MSC) block to replace some fixed-scale convolution layers, see Fig. 2. The MSC block begins with a MSC layer to extract multi-scale feature maps from input feature and then utilizes a 1×1 convolution layer to enhance the relation between multi-scale feature maps in channels, see Fig. 3a. For example, a MSC layer utilizes three different kernels to extract multi-scale information of the input feature and then concatenates the three output feature maps along the channel to get the output feature, see Fig. 3b. In practical application, considering the sizes of input

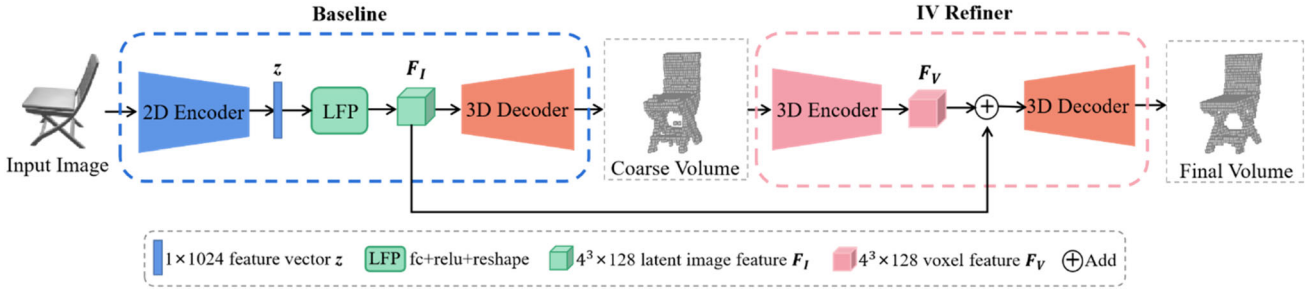


Fig. 1 Overview of the proposed IV-Net

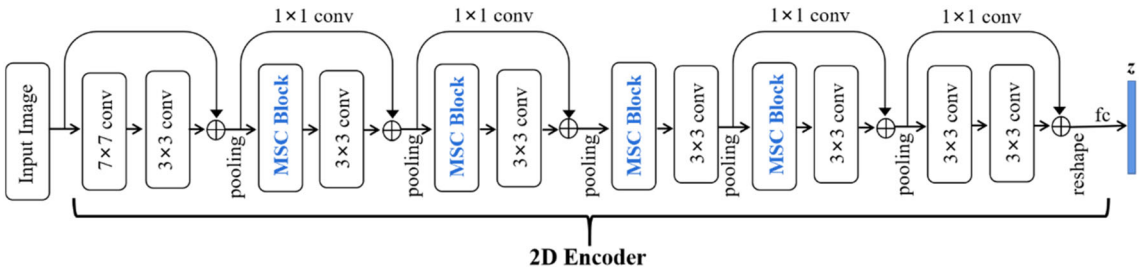
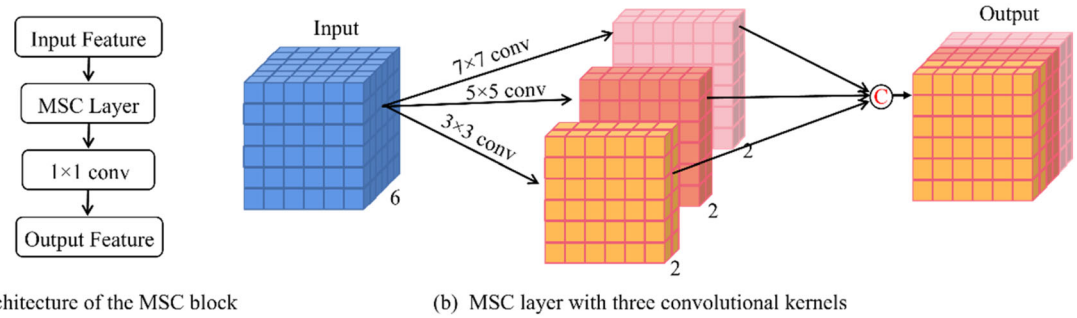


Fig. 2 2D encoder with MSC blocks, where fc is a fully connected layer



(a) Architecture of the MSC block (b) MSC layer with three convolutional kernels

Fig. 3 Multi-scale convolutional (MSC) block extracts the multi-scale information of input feature in 2D encoder. **a** Architecture of MSC block. In practical application, we only use two kernels (i.e., 3 × 3, 5 × 5) in each MSC layers. **b** MSC layer with three convolutional kernels ($c = 6$). 'C' denotes concatenating feature maps along the channel

image features, we only use two kernels (i.e., 3×3 conv, 5×5 conv) in MSC layers. More specifically, those convs in MSC layers are all with same stride 1×1, padding of 'SAME,' and filter c/n, where c is the number of channels of input feature, and n is the number of types of kernels.

3.1.2 Latent feature processing (LFP) module

LFP maps the 1×1024 feature vector $\mathbf{z}$ to a voxelized $4^3 \times 128$ latent image feature $\mathbf{F_I}$. AttSets [8] is a unified framework for single-view and multi-view 3D reconstruction and employs a LFP module to attentively fuse the image features and get a voxelized latent image feature. For a given single-view image, we can simplify the LFP, which only contains a fully connected (fc) layer, a relu activation and a reshape operation.

3.1.3 3D Decoder

3D decoder transforms latent feature $\mathbf{F_I}$ into a coarse volume of size $32 \times 32 \times 32$. And its construction is same with the 3D decoder of 3D-R2N2 [7].

3.1.4 IV Refiner

To supplement certain spatial information, based on pre-trained baseline, IV refiner fuses features of image and recovered coarse volume, then to get a more accuracy volume. It consists of two components: a 3D encoder and a 3D decoder. Following the structure of our 2D encoder, we construct 3D residual convolutional architectures for 3D encoder to effectively extract voxel feature.

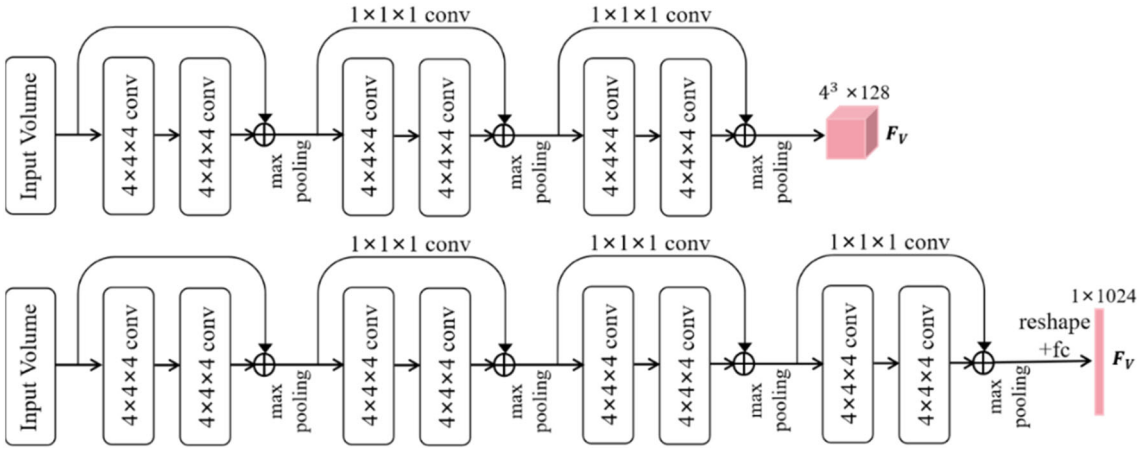


Fig. 4 Network architectures of 3D Encoder/A (top) and 3D Encoder/B (bottom)

3.1.5 3D Encoder

Generally, the combinations of features have two main methods: 'concat' and '+.' To get voxel feature $\mathbf{F_V}$ of the coarse volume, based on these two methods, we construct two versions of residual convolutional architectures for 3D encoder: 3D Encoder/A and 3D Encoder/B, as shown in Fig. 4. Every residual convolutional block begins with two banks of $4 \times 4 \times 4$ convolutional layers with stride $1 \times 1 \times 1$, where every layer is followed by a leaky relu activation with a leaky rate of 0.2, then adds the residual connection between input feature and the second layer, lastly follows a max pooling layer with kernel size of $2 \times 2 \times 2$. In 3D Encoder/A, there are three residual blocks, and the numbers of the output channels of convolutional layers in residual blocks are 32, 64 and 128, respectively. In 3D Encoder/B, it begins with four residual blocks and follows a reshape operation, a fc layer. The numbers of the output channels of convolutional layers in residual blocks are 32, 64, 128 and 128, respectively. After being processed by 3D Encoder/A or /B, the 3D volume input is encoded to a $4^3 \times 128$ voxel feature $\mathbf{F_V}$ or 1×1024 voxel feature $\mathbf{F_V}$, respectively.

3.1.6 3D Decoder

3D decoder of IV refiner is identical to that of baseline. It takes the feature combined by image feature and voxel feature $\mathbf{F_V}$ as input and then transforms the combined feature to generate a final volume. There are two main combination methods. '+' adds latent image feature $\mathbf{F_I}$ with $\mathbf{F_V}$, and 'concat' concatenates image feature $\mathbf{z}$ with $\mathbf{F_V}$ along the channel. Note that we adopt 3D encoder/A and /B while applying methods '+' and 'concat,' respectively.

3.2 Reconstruction loss

Reconstruction loss is crucial in network training. Suppose that Y denotes the ground truth, y represents the corresponding prediction, Y_i and y_i are the i-th ground truth voxel and predicted voxel, and N denotes the voxel number of the predicted volume. Two reconstruction losses are introduced as following.

3.2.1 Cross-entropy (CE) loss

The standard cross-entropy loss is always used as the loss function of previous works on 3D volume reconstruction [7–9, 11]. It is calculated as follows:

$$l_{\mathrm{CE}}(Y,\ y) = \frac{1}{N}\sum_{i=1}^{N}[Y_i \log(y_i) + (1 - Y_i)\log(1 - y_i)]. \tag{1}$$

3.2.2 Dice loss

The Dice can better optimize IoUs [21] and solve the problem of the highly unbalanced voxel occupancy [12, 22], and it is defined as follows:

$$l_{\mathrm{Dice}}(Y,\ y) = 1 - \frac{\sum_{i=1}^{N} Y_i y_i}{\sum_{i=1}^{N}(Y_i + y_i)} - \frac{\sum_{i=1}^{N}(1 - Y_i)(1 - y_i)}{\sum_{i=1}^{N}(2 - Y_i - y_i)}. \tag{2}$$

Usually, the smaller the value of lose function is, the closer the prediction is to the ground truth. Through comparative experiments, we finally choose Dice loss [22] as our reconstruction loss to better optimize baseline and IV refiner step by step.

4 Experiments

Our IV-Net is trained in Tensorflow 2.0 with an Intel Core i9-10920X CPU @ 3.50 GHz and a GeForce RTX 3060, and we set a batch size of 24 and adopt an Adam optimizer [23]. In this section, we show our experimental evaluations on two public datasets ShapeNet [24] and Pix3D [25]. For training and testing, the output 3D reconstructions are at size 32^3. For the training dataset, we first train the baseline module for 60 epochs, after freezing the pre-trained baseline, the IV refiner is trained for 40 epochs. In addition, we adopt Intersection over Union (IoU) and F-Score as the similarity evaluation metrics [26].

4.1 Datasets

4.1.1 ShapeNet

As a large 3D object dataset, ShapeNet [24] contains 55 categories and 51,300 3D models. Following [7–9, 13–17, 29, 30], as a subset of ShapeNet (i.e., ShapeNet13) is also utilized in our paper, which includes 44K models in the resolution of 32^3. For ShapeNet13, 24 images of size 137×137 for each model were rendered from 24 different viewpoints by 3D-R2N2 [7]. For our baseline module, the input size we need is 127×127. Hence, in our experiments, we just resize single-view images from 137×137 to 127×127.

4.1.2 Pix3D

Different from the synthetic dataset ShapeNet [24], Pix3D [25] aligns 3D models with real-world 2D images, and the largest category of it is the chair category, which consists of 3839 real-world images and corresponding objects. And according to the convention, the Pix3D is just used to evaluate the proposed methods in real-world images [9, 25]. Therefore, we also only test our proposed method on Pix3D-Chairs.

4.2 Evaluation metric

For the proposed networks, IoU is applied as a similarity metric to evaluate their reconstruction quality, and the IoU score is calculated as follows:

$$\mathrm{IoU} = \frac{\sum_{i=1}^{N} I(y_i > t) I(Y_i > 0)}{\sum_{i=1}^{N} I[(I(y_i > t) + I(Y_i > 0)) > 0]}, \tag{3}$$

where $I(\cdot)$ is an indicator function which will be 0 or 1 when the requirements are unsatisfied or satisfied, respectively, Y_i and y_i are the i-th ground truth voxel and predicted value, t is the threshold for voxelization which is setted [9] as a fixed value 0.3 in our experiments, N denotes the total voxel number of the predicted volume.

F-Score, as an extra metric, is also used to evaluate reconstruction quality of methods. And the F-Score is defined [27] as follows:

$$\text{F-Score}(d) = \frac{2P(d)R(d)}{P(d) + R(d)}, \tag{4}$$

where d is the distance threshold which is setted [9] as 1%, and $P(d)$ and $R(d)$ are the precision and recall. The precision $P(d)$ and recall $R(d)$ can be computed as follows:

$$P(d) = \frac{1}{N_R} \sum_{r \in R} \left[\min_{g \in G} \|g - r\| < d \right], \tag{5}$$

$$R(d) = \frac{1}{N_G} \sum_{g \in G} \left[\min_{r \in R} \|g - r\| < d \right], \tag{6}$$

where R and G present the predicted and ground truth point clouds, andN_R and N_G denote the total number of points in the R and G. For voxel-based reconstruction methods, we first generate mesh of 3D surface from voxel by applying marching cubes algorithm [28] and then sample 8192 points [9] from mesh to obtain corresponding point cloud.

4.3 Ablation study

In this section, IV-Net is ablated by utilizing simplified LFP, loss functions, MSC block and 3D encoders of IV refiner on the ShapeNet dataset [24]:

Table 1 The effect of the Dice loss, MSC block, and IV refiner in our proposed network in terms of IoU and F-Scores

	Loss function	Simplified LFP	MSC block	IV refiner	IoU	F-scores
AttSets	CE				0.642	0.395
AttSets/S	CE	√			0.642	0.395
AttSets/S	Dice	√			0.655	0.416
Baseline	Dice	√	√		0.658	0.421
Baseline	Dice	√	√	'Concat'	0.680	0.440
IV-Net	Dice	√	√	'+'	0.681	0.443

Table 2 Comparisons of parameter size of the two methods 'concat' and '+' to fuse image feature and corresponding voxel feature

Method	' + '	'Concat'
#Parameters	23.63 M	50.58 M

Set up 1 To validate the rationality of simplifying LFP, we train AttSets [8] with original or simplified LFP with standard CE loss, respectively. Table 1 shows that we can maintain learning effect of LFP while simplifying it. We define AttSets with simplified LFP as AttSets/S.

Set up 2 To compare what losses benefit the performance of our method, the AttSets/S is also trained with Dice loss [22]. Table 1 indicates that using Dice loss to replace CE loss [7] causes performance upgradation from 0.642 to 0.655. Therefore, Dice loss is optimal to as our reconstruction loss.

Set up 3 One might think that the fixed 3×3 convolution limits the ability to extract image features. And we replace some fixed 3×3 convolution layers with MSC blocks. We can see that the performance increases from 0.651 to 0.654 after using MSC blocks. Trained with reconstruction loss l_{Dice}, AttSets/S with MSC blocks is defined as our baseline.

Set up 4 A key issue of IV-Net is how to fuse features of image and recovered volume. We compare two most commonly methods '+' and 'concat' in Table 1. And 3D encoder/A and /B are, respectively, adopted while applying methods '+' and 'concat.' Table 1 shows the performance of IV refiner using method '+' is better than the 'concat,' and adding IV refiner improves the performance from 0.658 to 0.681 or 0.680. Table 2 indicates the method '+' has less parameters than 'concat.' Hence, IV refiner utilizes '+' to fuse features and chooses 3D encoder/A as shape encoder.

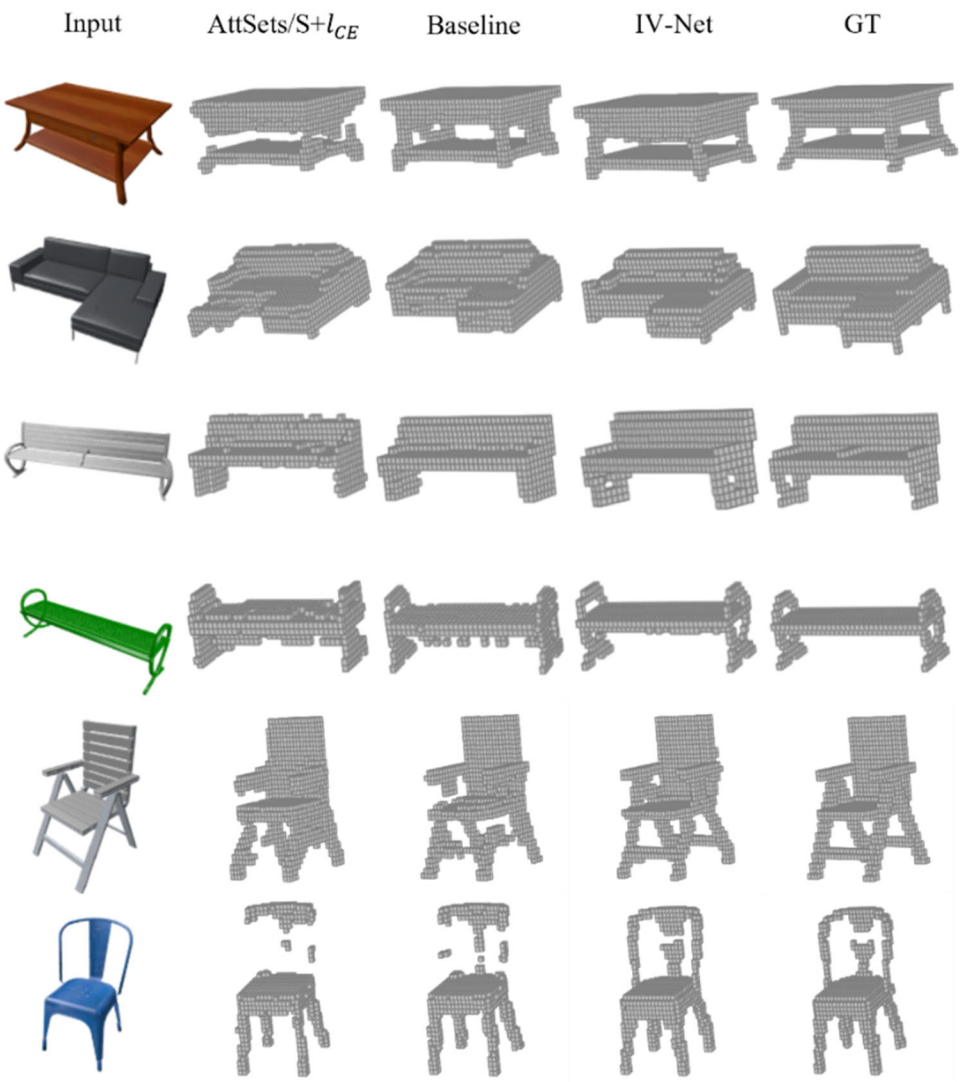


Fig. 5 Visual comparisons of AttSets/S, baseline and IV-Net

Moreover, we also give some visual comparisons of AttSets/S with CE loss, baseline and IV-Net ('+') in Fig. 5. And Fig. 5 indicates that our baseline reaches better performance than AttSets/S with CE loss and IV-Net outperforms baseline, which validate the effect of Dice loss, MSC block and IV refiner using 3D encoder/A.

4.4 Evaluation on the ShapeNet dataset

On the synthetic ShapeNet dataset [24], we split ShapeNet into two sets, with 4/5 of it to train and the remaining to test, same as [7, 8]. IV-Net is compared with several state-of-the-art methods, containing 3D-R2N2 [7], OGN [13], Matryoshka [14], AtlasNet [29], Pixel2Mesh [15], OccNet [30], IM-Net [31], AttSets [8], and Pix2Vox++ [9], and the IoU scores and F-Scores of these methods are illustrated in Tables 3 and 4, respectively, where the overall IoU/F-Score are taken as the mean IoU/F-Score across all 13 categories. For the overall IoU and F-Score, we observe that our IV-Net outperforms these methods. Additionally, IV-Net outperforms all other methods in 5 of the 13 categories about IoU and in 4 of 13 about F-Score.

Table 3 IoU results of several reconstruction approaches on ShapeNet13. For each category, the best IoU score is highlighted in bold

Category	3D-R2N2	OGN	Matryoshka	AtlasNet	Pixel2Mesh	OccNet	IM-Net	AttSets	Pix2Vox++	IV-Net
Airplane	0.513	0.587	0.647	0.493	0.508	0.532	**0.702**	0.594	0.674	0.701
Display	0.468	0.502	0.532	0.457	0.582	**0.651**	0.585	0.565	0.548	0.614
Telephone	0.661	0.702	0.756	0.543	0.762	0.794	0.762	0.743	**0.809**	0.792
Watercraft	0.560	0.632	0.591	0.355	0.471	0.579	0.607	0.601	0.603	**0.630**
Speaker	0.662	0.637	0.701	0.296	0.672	0.655	0.683	0.721	0.721	**0.723**
Lamp	0.381	0.398	0.408	0.261	0.399	0.474	0.433	0.445	0.457	**0.487**
Bench	0.421	0.481	0.577	0.431	0.379	0.597	0.564	0.552	0.608	**0.611**
Rifle	0.544	0.593	0.616	0.573	0.468	0.656	**0.723**	0.601	0.617	0.672
Sofa	0.628	0.646	0.681	0.354	0.622	0.669	0.694	0.703	0.725	**0.737**
Cabinet	0.716	0.729	0.776	0.257	0.732	0.674	0.680	0.783	**0.799**	0.796
Car	0.798	0.828	0.850	0.282	0.670	0.671	0.756	0.844	**0.858**	0.856
Chair	0.466	0.483	0.547	0.328	0.484	0.583	**0.644**	0.559	0.581	0.597
Table	0.513	0.536	0.573	0.301	0.536	**0.659**	0.621	0.590	0.620	0.635
Overall	0.560	0.596	0.635	0.352	0.552	0.626	0.659	0.642	0.670	**0.681**

Table 4 F-Score results of several reconstruction methods on ShapeNet13. For each category, the best IoU score is highlighted in bold

Category	3D-R2N2	OGN	Matryoshka	AtlasNet	Pixel2Mesh	OccNet	IM-Net	AttSets	Pix2Vox++	IV-Net
Airplane	0.412	0.487	0.446	0.415	0.376	0.494	**0.598**	0.489	0.583	0.594
Display	0.227	0.215	0.400	0.451	0.319	**0.468**	0.466	0.310	0.296	0.387
Telephone	0.504	0.528	0.598	0.545	0.485	0.273	0.423	0.469	**0.633**	0.584
Watercraft	0.305	0.328	0.360	0.296	0.266	0.347	0.369	0.315	**0.390**	0.370
Speaker	0.231	0.225	**0.279**	0.199	0.190	0.249	0.200	0.211	0.152	0.216
Lamp	0.267	0.249	0.276	0.217	0.219	0.361	0.371	0.315	0.315	**0.373**
Bench	0.345	0.364	0.424	0.439	0.313	0.318	0.361	0.406	0.478	**0.480**
Rifle	0.521	0.541	0.514	0.405	0.340	0.219	0.407	0.524	0.574	**0.578**
Sofa	0.274	0.290	0.326	0.337	0.343	0.324	0.354	0.334	0.377	**0.379**
Cabinet	0.327	0.316	0.381	0.350	**0.450**	0.449	0.345	0.367	0.408	0.391
Car	0.481	0.514	0.481	0.319	0.486	0.315	0.304	0.497	**0.564**	0.549
Chair	0.238	0.226	0.302	0.406	0.386	0.365	**0.442**	0.334	0.309	0.402
Table	0.340	0.352	0.374	0.371	0.502	**0.549**	0.461	0.419	0.406	0.458
Overall	0.351	0.368	0.391	0.362	0.398	0.393	0.405	0.395	0.436	**0.443**

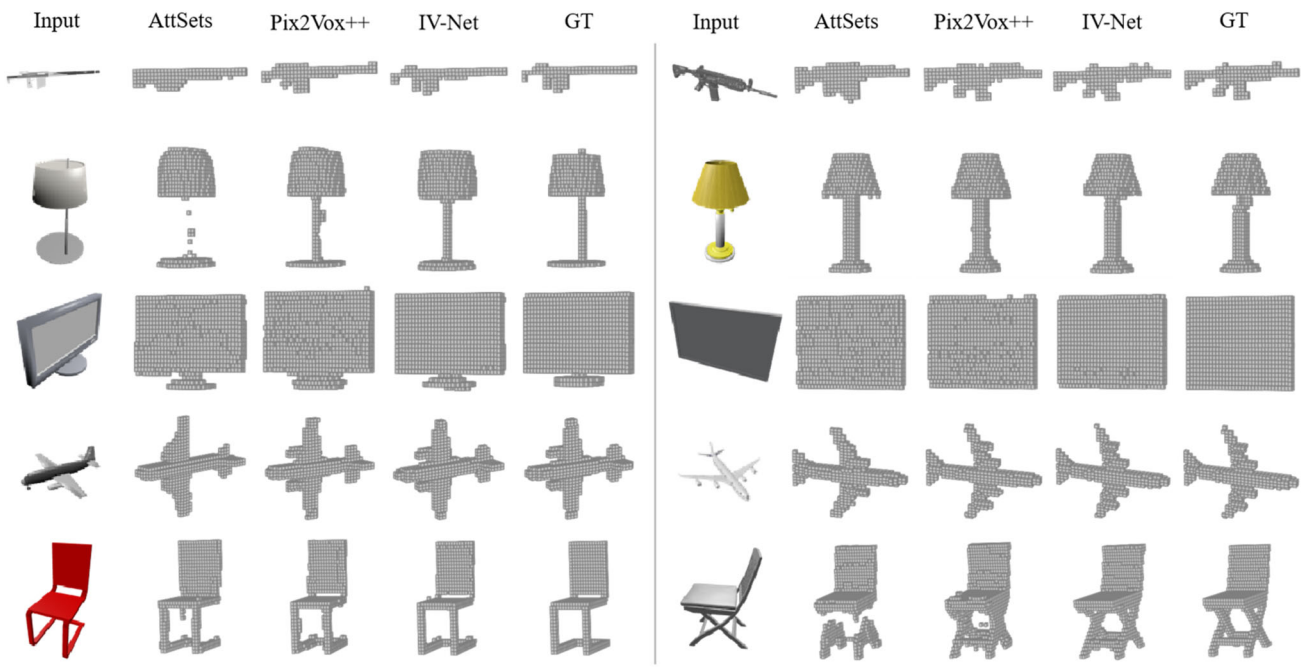


Fig. 6 Visual examples of single-view 3D reconstruction on ShapeNet13

Table 5 IoU and F-Score results of several reconstruction approaches on Pix3D-Chairs

Method	3D-R2N2	Pix3D	Pix2Vox++	IV-Net
IoU	0.136	0.282	0.288	0.292
F-Scores	0.018	0.041	0.068	0.109

Meanwhile, in visual effect, IV-Net is compared with two voxel-based approaches AttSets [8] and Pix2Vox++ [9] in Fig. 6, which indicates that IV-Net reconstructs more visually cleaner and accurate volumes in some categories. For instance, IV-Net shows more accurate reconstruction results than AttSets and Pix2Vox++ in the legs of chairs, the tail and wings of airplanes, small details in rifles and lamps, and so on.

4.5 Evaluation on the Pix3D dataset

On the real-world Pix3D dataset [25], following [9, 25], we also use Pix3D-Chairs as the testing set, to evaluate methods on real-world images. Considering the complex backgrounds of real-world images, using Render for CNN [32], we need first generate 60 images for each chair of ShapeNet-Chairs by adding random backgrounds [9, 25], sampled from the dataset SUN [33]. And these generated images are used as the training set, i.e., ShapeNet-Chairs-RfS. Our IV-Net is compared with 3D-R2N2 [7], Pix3D [25], and Pix2Vox++ [9]. The IoUs and F-Scores on Pix3D-Chairs are shown in Table 5, and the results indicate that our IV-Net performs better than these methods. Figure 7 gives some visual comparisons on Pix3d-Chairs among our baseline, Pix2vox++ and IV-Net. Through adding additional spatial feature, IV-Net obtains better reconstruction than the baseline and Pix2vox++ on the details of objects, such as legs and handles.

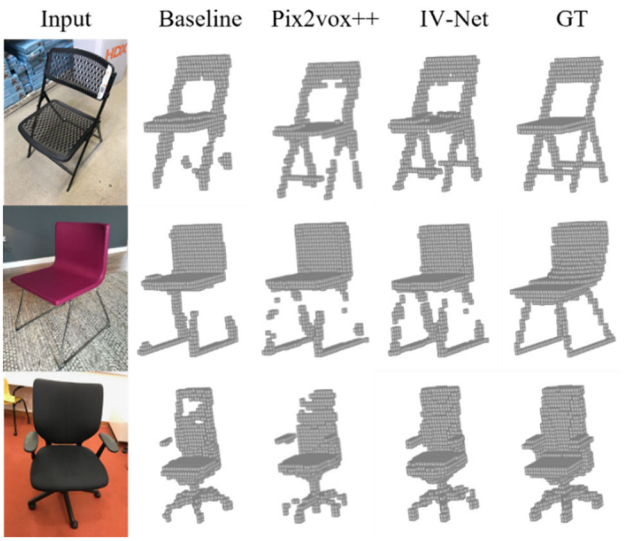


Fig. 7 Examples of single-view 3D reconstruction on Pix3D-Chairs

Table 6 Parameter size and inference time comparisons among IV-Net and other state-of-the-art methods

Method	3D-R2N2	AtlasNet	Pixel2Mesh	IM-Net	AttSets	Pix2Vox++	IV-Net
#Parameters (M)	36	45	21	55	18	98	24
Inference time (ms)	78.86	38.47	60.78	10.89	26.32	10.64	27.13

4.6 Computational complexity

For computational complexity of different methods, the parameter size and inference time of IV-Net and some state-of-the-art methods are compared in Table 6. The values of Table 6 are collected from Pix2Vox++ [9], and we follow its scheme to get the values of our method.

5 Conclusions

In this paper, we propose a novel framework for single-view 3D reconstruction, named IV-Net, which has universal and adaptable application prospect. In our proposed method, we design multi-scale convolutional block to enhance the ability of 2D encoder and construct two versions of 3D encoders to extract voxel feature efficiently. By fusing features of image and recovered volume, an IV refiner raises the accuracy of the reconstructed volumes and recovers the detailed structures of 3D shapes. In both quantitative and qualitative evaluations, our network outperforms state-of-the-art methods in 3D reconstruction and has less parameters than mostly methods. However, our proposed method does not obtain the optimal results on some categories of ShapeNet13. In future, we will continue to make our network better.

Funding This work was supported by National Natural Science Foundation of China (Grant Number [11471093]).

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that there are no conflicts of interest.

References

1. Montefusco, L.B., Lazzaro, D., Papi, S., et al.: A fast compressed sensing approach to 3D MR image reconstruction. IEEE Trans. Med. Imaging **30**(5), 1064–1075 (2010). https://doi.org/10.1109/TMI.2010.2068306
2. Sra, M., Garrido-Jurado, S., Schmandt, C., Maes, P.: Procedurally generated virtual reality from 3D reconstructed physical space. In: ACM Conference on Virtual Reality Software and Technology, pp. 191–200 (2016). https://doi.org/10.1145/2993369.2993372
3. Avetisyan, A., Dahnert, M., Dai, A., et al.: Scan2CAD: learning CAD model alignment in RGB-D scans. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2609–2618 (2019). https://doi.org/10.1109/CVPR.2019.00272
4. Popa, A.I., Zanfir, M., Sminchisescu, C.: Deep multitask architecture for integrated 2d and 3d human sensing. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4714–4723 (2017). https://doi.org/10.1109/CVPR.2017.501
5. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge (2003)
6. Durrant-Whyte, H., Bailey, T.: Simultaneous localization and mapping: part I. IEEE Robot. Autom. Mag. **13**(2), 99–110 (2006)
7. Choy, C.B, Xu, D., Gwak, J.Y., et al.: 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In: European Conference on Computer Vision, pp. 628–644. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_38
8. Yang, B., Wang, S., Markham, A., et al.: Robust attentional aggregation of deep feature sets for multi-view 3D reconstruction. Int. J. Comput. Vis. **128**(1), 53–73 (2020). https://doi.org/10.1007/s11263-019-01217-w
9. Xie, H., Yao, H., Zhang, S., Zhou, S., Sun, W.: Pix2Vox++: multi-scale context-aware 3D object reconstruction from single and multiple images. Int. J. Comput. Vis. **128**(12), 2919–2935 (2020). https://doi.org/10.1007/s11263-020-01347-6
10. Wu, J., Zhang, C., Xue, T., et al.: Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. Adv. Neural Inf. Process. Syst. **29** (2016)
11. Yagubbayli, F., Tonioni, A., Tombari, F.: Legoformer: transformers for block-by-block multi-view 3D reconstruction. arXiv preprint arXiv:2106.12102 (2021)
12. Shi, Z., Meng, Z., Xing, Y., et al.: 3D-RETR: end-to-end single and multi-view 3D reconstruction with transformers. arXiv preprint arXiv:2110.08861 (2021)
13. Tatarchenko, M., Dosovitskiy, A., Brox T.: Octree generating networks: efficient convolutional architectures for high-resolution 3d outputs. In: IEEE International Conference on Computer Vision, pp. 2107–2115 (2017). https://doi.org/10.1109/ICCV.2017.230
14. Richter, S.R., Roth, S.: Matryoshka networks: predicting 3D geometry via nested shape layers. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1936–1944 (2018). https://doi.org/10.1109/CVPR.2018.00207
15. Wang, N., Zhang, Y., Li, Z., et al.: Pixel2mesh: generating 3D mesh models from single RGB images. In: European Conference on Computer Vision, pp. 52–67 (2018)
16. Fan, H., Su, H., Guibas, L.J..: A point set generation network for 3D object reconstruction from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2463–2471 (2017). https://doi.org/10.1109/CVPR.2017.264
17. Mandikal, P., Navaneet, K.L., Agarwal, M., et al.: 3D-LMNet: Latent embedding matching for accurate and diverse 3D point cloud reconstruction from a single image. arXiv preprint arXiv:1807.07796 (2018)

18. Nozawa, N., Shum, H.P.H., Feng, Q., et al.: 3D car shape reconstruction from a contour sketch using GAN and lazy learning. Vis. Comput. **38**, 1317–1330 (2022). https://doi.org/10.1007/s00371-020-02024-y
19. Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al.: Generative adversarial nets. Adv. Neural Inf. Process. Syst. (2014)
20. Wiles, O., Zisserman, A.: SilNet: single- and multi-view reconstruction by learning from silhouettes. In: British Machine Vision Conference (2017)
21. Berman, M., Triki, A.R., Blaschko, M.B.: The lovasz-softmax loss: a tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4413–4421 (2018). https://doi.org/10.1109/CVPR.2018.00464
22. Sudre, C.H., Li, W., Vercauteren, T., et al.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 240–248. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_28
23. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
24. Chang, A.X., Funkhouser, T., Guibas, L., et al.: Shapenet: an information-rich 3D model repository. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1912–1920 (2015)
25. Sun, X., Wu, J., Zhang, X., et al.: Pix3d: dataset and methods for single-image 3D shape modeling. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2974–2983 (2018). https://doi.org/10.1109/CVPR.2018.00314
26. Li, Y., Wang, Z., Yin, L., et al.: X-Net: a dual encoding–decoding method in medical image segmentation. Vis. Comput. (2021). https://doi.org/10.1007/s00371-021-02328-7
27. Tatarchenko, M., Richter, S.R., Ranftl, R., Li, Z., Koltun, V., Brox, T.: What do single-view 3D reconstruction networks learn? In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3400–3409 (2019). https://doi.org/10.1109/CVPR.2019.00352
28. Lorensen, W.E., Cline, H.E.: Marching cubes: a high resolution 3D surface construction algorithm. ACM Siggraph Comput. Graph. **21**(4), 163–169 (1987). https://doi.org/10.1145/37402.37422
29. Groueix, T., Fisher, M., Kim, V.G., et al.: AtlasNet: a Papier–Mâché approach to learning 3D surface generation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 216–224 (2018)
30. Mescheder, L., Oechsle, M., Niemeyer, M., et al.: Occupancy networks: learning 3D reconstruction in function space. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4455–4465 (2019). https://doi.org/10.1109/CVPR.2019.00459
31. Chen, Z., Zhang, H.: Learning implicit fields for generative shape modeling. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5932–5941 (2019). https://doi.org/10.1109/CVPR.2019.00609
32. Su, H., Qi, C.R., Li, Y., et al.: Render for CNN: viewpoint estimation in images using CNNs trained with rendered 3D model views. In: IEEE International Conference on Computer Vision, pp. 2686–2694 (2015). https://doi.org/10.1109/ICCV.2015.308
33. Xiao, J., Hays, J., Ehinger, K.A., et al.: 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

Beibei Sun received B.S. degree in school of Mathematics, Hefei Normal University, Anhui, China, in 2017. She is currently a master's degree student in the School of Mathematics, Hefei University of Technology, Anhui, China. Her research interests include image processing, deep learning, computer vision, 3D geometry processing and 3D reconstruction.

Ping Jiang received B.S. degree in school of Mathematics, East China Normal University, Shanghai, China, in 1995. She received Ph.D. degree in school of Computer Science, Hefei University of Technology, Anhui, China, in 2006. She is a professor in the School of Mathematics, Hefei University of Technology, Anhui, China. Her research interests include approximation theory, CAGD, computer vision and deep learning.

Dali Kong received B.S. degree in school of Mathematics, Hefei University, Anhui, China, in 2018. He is currently a master's degree student in the School of Mathematics, Hefei University of Technology, Anhui, China. His research interests include image processing, deep learning, computer vision, 3D geometry processing and point cloud denoising.

Ting Shen received B.S. degree in school of Mathematics, Hefei Normal University, Anhui, China, in 2021. She is currently a master's degree student in the School of Mathematics, Hefei University of Technology, Anhui, China. Her research interest is approximation theory, CAGD, Approximation theory, computer vision, deep learning and image processing.