Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(8): 13864–13880.
DOI: 10.3934/mbe.2023617
Received: 30 November 2022
Revised: 14 May 2023
Accepted: 21 May 2023
Published: 19 June 2023

Research article

IFKD: Implicit field knowledge distillation for single view reconstruction

Jianyuan Wang[1,2], Huanqiang Xu[3], Xinrui Hu[3] and Biao Leng[3,*]

[1] School of Intelligence Science and Technology, University of Science and Technology Beijing, Beijing 100083, China

[2] Key Laboratory of Intelligent Bionic Unmanned Systems, Ministry of Education, University of Science and Technology Beijing, Beijing 100083, China

[3] School of Computer Science and Engineering, Beihang University, Beijing 100191, China

* **Correspondence:** Email: lengbiao@buaa.edu.cn.

Abstract: In 3D reconstruction tasks, camera parameter matrix estimation is usually used to present the single view of an object, which is not necessary when mapping the 3D point to 2D image. The single view reconstruction task should care more about the quality of reconstruction instead of the alignment. So in this paper, we propose an implicit field knowledge distillation model (IFKD) to reconstruct 3D objects from the single view. Transformations are performed on 3D points instead of the camera and keep the camera coordinate identified with the world coordinate, so that the extrinsic matrix can be omitted. Besides, a knowledge distillation structure from 3D voxel to the feature vector is established to further refine the feature description of 3D objects. Thus, the details of a 3D model can be better captured by the proposed model. This paper adopts ShapeNet Core dataset to verify the effectiveness of the IFKD model. Experiments show that IFKD has strong advantages in IOU and other core indicators compared with the camera matrix estimation methods, which verifies the feasibility of the new proposed mapping method.

Keywords: single view reconstruction; implicit field; knowledge distillation; encoder-decoder

1. Introduction

3D reconstruction [1, 2] is one of the frontier research directions of computer graphics, which is widely used in virtual reality, medical treatment, architecture, industrial design, 3D printing, and many other fields. Traditional 3D modeling needs a lot of work of professionals. Although with the development of technology, users can obtain 3D objects through acquisition devices, even the cheapest depth cameras are much more expensive than ordinary cameras. Professional 3D acquisition devices are so expensive and not suitable for large-scale application. In recent years, with the

development of artificial intelligence and deep learning, the ability to perceive 3D models has been significantly improved. As a result, the demand for sensing equipment capability reduced a lot, which further promoted wider application of 3D reconstruction technology.

Single view reconstruction (SVR) [3–7] can reduce the demand for 3D object information collection, and build the whole shape of 3D objects in a delicate way. Because of these advantages, SVR has gradually become one of the mainstream methods in 3D reconstruction. SVR mainly relies on the encoder decoder structure to extract features from the input single image, and then generates a 3D model through restoration according to the features. Implicit Field model is one of the representative models used to solve SVR task by defining continuous functions in 2D/3D space, finding the zero isosurface of the field to reconstruct the mesh surface.

In most 3D reconstruction models, the estimation of camera parameters is an important step in object feature extraction. However, in the task of single view reconstruction, the parameter estimation of the camera itself is not necessary. It can even get rid of the dependence on the camera parameter matrix by assuming the initial position state of the object, so as to reduce the redundant structure in the 3D reconstruction model. In this way, the model itself can pay more attention to the reconstruction of 3D objects details, rather than aligning the objects framework. Instead of simplifying the feature extraction network structure, the proposed model omits the camera parameter matrix prediction network. Previous algorithms need to use camera parameters to map 3D points to image points before feature extraction. However, the proposed model does not require the camera matrix, which simplifies the mapping process rather than the network structure. The advantage of this is that no redundant network is required to predict the camera parameter matrix.

In this paper, we propose a Implicit Field Knowledge Distillation model (IFKD) for SVR task in Figure 1, which can deal with the task when the Camera Matrix is unknown. A new mapping method from 3D points to 2D pixels is proposed, instead of the estimation of Camera Matrix. In order to get more refined feature representation of the object, we adopt a knowledge distillation [8] structure to teach the feature extraction networks by voxel 3D encoder. The 3D voxel network can easily extract spatial information, so using the 3D voxel network to supervise the 2D student network can help the student networks learn spatial information better.

The main contributions of this paper are summarized as follows:

1) A new mapping method is proposed to reconstruct 3D objects without camera matrix, simplifying the network structure of feature extraction.

2) A teacher-student structure for voxel 3D feature knowledge distillation is adopted to refine the details of 3D objects.

3) The skip connection of the encoder-decoder is considered and discussed in detail to further optimize the effect of SVR.

The rest of the paper is organized as follows: Section 2 investigates the recent research progress of related work. Section 3 provides the pipeline and structures of implicit field knowledge distillation model. In Section 4, ShapeNet dataset is introduced to verify the proposed model, and experimental comparisons and discussions are provided. At last, Section 5 presents the conclusion.

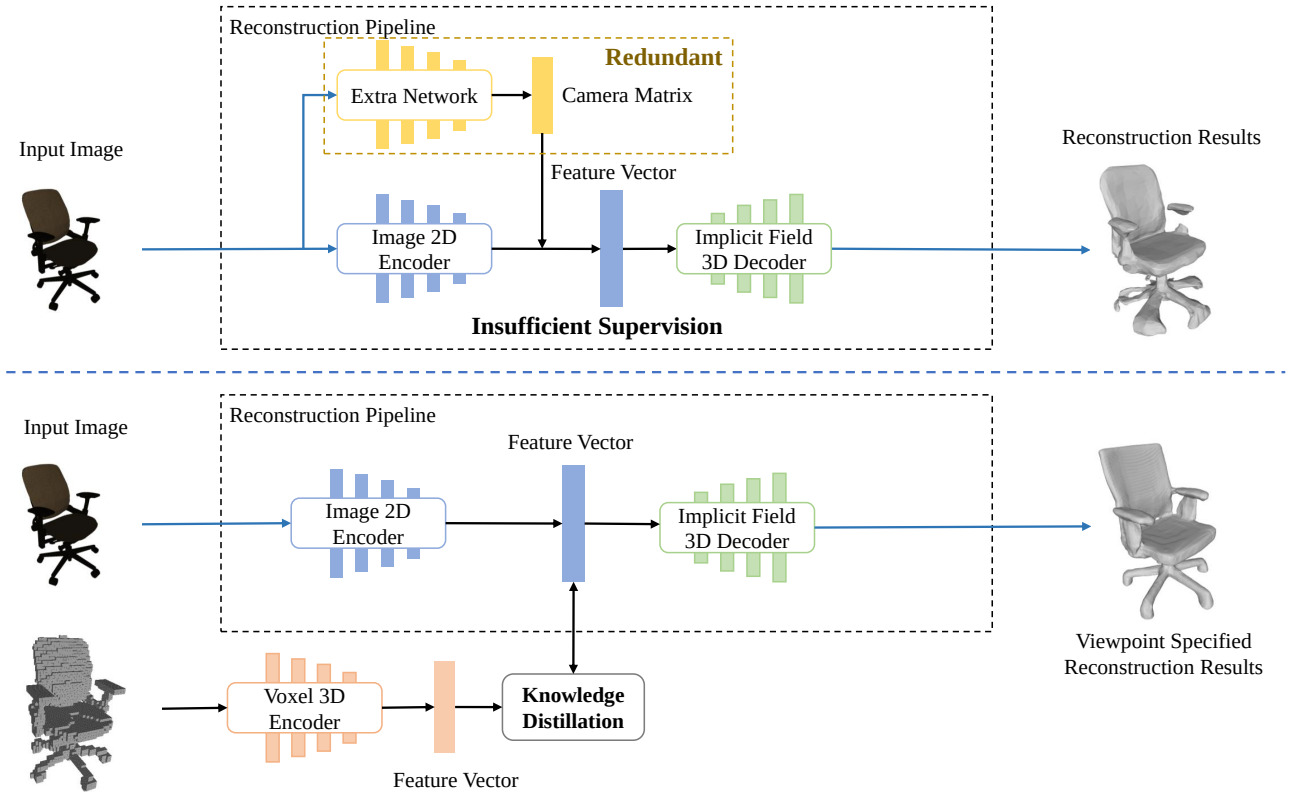

Figure 1. Motivation of the IFKD model. The model adopts a knowledge distillation model to refine the feature vector instead of a camera matrix estimation network.

2. Related works

2.1. Implicit-field-based SVR methods

According to the 3D representation of the reconstruction results, the existing SVR methods can be divided into Euclidean representation, non-Euclidean representation and implicit field based representation. The first two can be collectively classified as geometry-based methods. Models reconstructed by geometry-based methods can be represented by meshes [3, 9], voxels [4, 10, 11] and points [5, 12], they are intuitive and visible. However, it is difficult to analyze meshed points because they are sparse and irregular. Meanwhile, the storage occupied by a voxel grows cubically with its resolution, therefore, it is difficult to balance reconstruction quality with storage cost.

However, Implicit field learning avoids the limitation of storage cost and has achieved significant improvement in SVR. [6, 7, 13] use a similar idea of making predictions for each 3D point to reconstruct 3D shapes. This design is beneficial for generating contiguous surfaces because it does not use complex geometric representations and avoids the storage cost constraints. Meanwhile, it can encode descriptions of 3D output at infinite resolution without taking up excessive memory, which makes implicit field learning have greater advantages in the quality and the storage cost of the output voxels. However, the above models have insufficient perception of objects, resulting in insufficient accuracy of the reconstructed shapes [14]. The details of the objects, such as edges and corners, are

disconnected or linked up at the wrong scale. To address this problem, [1] proposes D2IM-Net to encode the input image as global and local features, which are fed into two decoders respectively. The base decoder uses global features to reconstruct the coarse implicit field, while the detail decoder reconstructs two displacement maps from local features. The final 3D reconstruction result is obtained by combining the base shape and displacement map. [15] predicts the projected position of each 3D point on a 2D image and extracts local features from image feature maps. Combining global and local features significantly improves the accuracy of signed distance field predictions, especially for detail-rich regions. Furthermore, [16] extracts local information from different layers and reorganizes this information.

2.2. Detail reconstruction

SVR has less input images and thus less useful information can be obtained. This results in that the results of SVR often lack detailed information. Therefore, researchers aim to solve the problem of insufficient detailed information in SVR. The current mainstream method is to use local feature encoding to solve this problem. In [1], the 3D shape reconstruction task is decomposed into two parts: shape reconstruction and residual reconstruction. The former aims to generate the main shape of the model using global feature vectors, while the latter focuses on reconstructing the details of the model. At the same time, the reconstructed result is projected to a 2D plane and compared with the original image, and the difference between these two images will be used as part of the loss function. [2] introduces an implicit representation function that aligns 2D image pixels with the global information of their corresponding 3D objects, which enables the function to infer the surface texture of the 3D reconstructed model using a single input image or multiple input images information if available.

Besides local feature encoding, inspired by how humans learn, [17] proposes a minimum circumference loss that trains the network in an easy-to-hard way. In the early stage of training, the network learns to reconstruct the main body through a high loss function tolerance. After that, the penalty for false prediction is increased to supervise the model to learn the details of the model.

2.3. Knowledge distillation

Knowledge Distillation [8] is a method of compressing knowledge from the cumbersome model into a more easily deployable model. In the pioneering work of classification, [8] proposes the teacher-student framework where the student network mimics the softened output of the teacher network. FitNet [18] extends this idea by training a student that is deeper and thinner than the teacher, using the output and intermediate representations learned by the teacher as cues to improve the training process and the performance of the student.

In addition, the teacher-student framework can also be used in object detection tasks. [19] make the student sample from the entire feature map, and then use a transformation layer to map the student's sampling results to the same dimensions as the teacher's sampling features. When training the student, [19] supervises the student to learn from the teacher by optimizing the similarity of the same area of the feature map sampled by the both networks. However, this method does not work on detectors without proposals. [20] supervises the student to learn the teacher's method of feature extraction and generalization based on the fact that the detector is more concerned with local near object region, but the introduction of the additional selection algorithm will increase the complexity

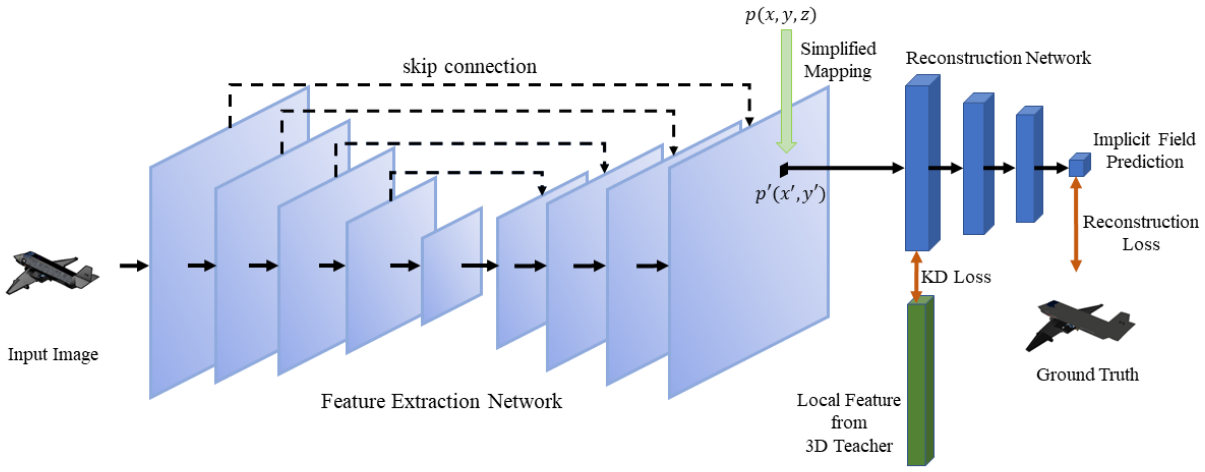

Figure 2. Pipeline of IFKD model. Local feature from 3D teacher is adopted to refine the reconstruction network.

of the network. [6] first employs a cue-based learning approach that encourages the feature representation of the student to be consistent with the teacher. After that, knowledge distillation is used to learn a stronger classification module. However, the imitation of the teacher's feature representation by the student will make the student network have a large amount of irrelevant noise, which makes the performance of the student not outstanding.

3. Method

3.1. Mapping 3D points to 2D pixels

The previous methods [1, 2, 15] have proven that local features from 2D images are important for improving the reconstruction of details. Extracting local features of a 3D point from a 2D image need map the point to a pixel first. The mapping from 3D point $p(x, y, z)$ to 2D pixel $p'(x', y')$ can be represented as the Camera Matrix M.

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = M \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{3.1}$$

However, the camera matrix is commonly unknown in our SVR settings. The existing methods train an extra CNN to predict the matrix, and then use it as the 3D to 2D mapping. This design introduces redundant network and accumulative errors.

The camera matrix M can be deduced from an intrinsic matrix K and an extrinsic matrix E,

formulated as follows,

$$M = KE = \begin{bmatrix} f/d_x & 0 & u_0 \\ 0 & f/d_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} \tag{3.2}$$

where intrinsic matrix K is determined by the properties of the camera itself, f is focal length of camera. d_x and d_y represent the scale relationship between the camera coordinate and the camera coordinate, while u_0 and v_0 represent the offset between the origins of these two coordinates. Extrinsic matrix E indicates the transform from the world coordinate to camera coordinate, where R and t represent the rotation and translation transformations, respectively.

Given a 3D object, different images can be captured by the same camera but with different positions and rotations. Each image I_i is corresponding to a camera matrix M_i, which consists of a constant intrinsic matrix K and a variable extrinsic matrix E_i. By adjusting properties of the camera, we can assume that the offset represented by u_0 and v_0 in K is zero and $d = d_x = d_y$. So the intrinsic matrix K is simplified as follows,

$$K = \begin{bmatrix} f/d & 0 & 0 \\ 0 & f/d & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.3}$$

Further, we can perform transformations on 3D points instead of the camera and keep the camera coordinate identified with the world coordinate, so that the extrinsic matrix can be omitted.

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = KE \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = M_c(E \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}) \tag{3.4}$$

where M_c is the simplified camera matrix. Now all images are corresponding to the same camera matrix M_c, but different transformation E_i on the 3D object. In fact, M_c is just depended on the resolution ratio of the 2D image and the 3D space, and can be fixed.

As the same with existing implicit-field based methods, the training samples of our method are point-value pairs (p, v)s, where p is 3D point and v indicate whether p is occupied by the 3D object. In training phrase, the transformation E_i of image I_i is known. So we can perform E_i on point p, when the network input is I_i. The transformed point p can be mapped to a 2D point p' directly, then the local feature can be extracted from the CNN feature map as described in 3.2.

In inference phrase, E_i is unknown, but also unnecessary. We directly feed the image I_i and the points into network, get the implicit field and then generate surfaces by the Marching Cubes algorithm [21]. Without transformation on input points, the reconstructed result is not aligned and has the same orientation with the object in I_i. This is different with existing methods that pursue aligned results and the consistent orientation. We argue that the SVR task should care more about the quality of reconstruction instead of the alignment.

3.2. Network architecture

Our network consists of a feature extraction network and a reconstruction network. The backbone of feature extraction network is a U-Net [22], which is commonly used in semantic segmentation.

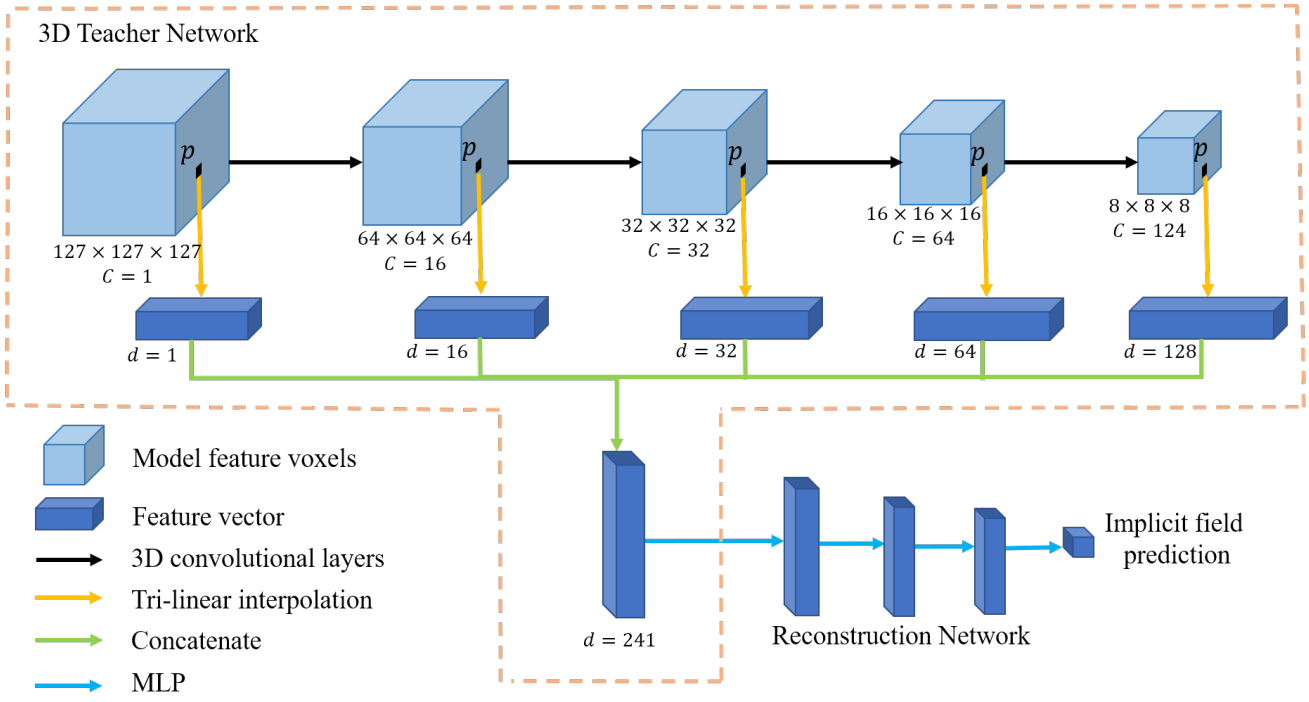

Figure 3. Teacher network structure for feature vector knowledge distillation.

The network utilizes skip connections between shallow layers and deep layers to reserve more texture information. The reconstruction network is a simple Multi-layer Perceptron(MLP) [23].

The U-Net takes image I as input and outputs a feature map. To predict the implicit field of point p, we first map it to 2D point p' on the feature map as described in 3.1. Then we utilize bi-linear interpolation algorithm to extract the feature vector of p. The vector is fed into reconstruction network, and the network outputs implicit field prediction.

3.3. Knowledge distillation from 3D

As shown in Figure 2, our SVR network is supervised by two losses. The first one is reconstruction loss, which is the same with existing methods. The second one is Knowledge Distillation (KD) loss. We use a 3D network as the teacher and our SVR network as the student. Then, the KD loss is designed to force the student to learn from the teacher network.

To obtain the teacher network, we first train a voxel-to-implicit field network, as shown in Figure 3. The encoder is a 3D CNN, which processes the voxels of object and output a feature voxels. Similar to the SVR pipeline, we use tri-linear interpolation to extract the local feature vector of 3D points from the feature voxels. Then the feature vector is fed into the decoder to get the implicit field prediction. The decoder is implemented with MLP. The teacher network is trained to transform the representation of 3D objects. At the same time, the teacher network learns to extract local feature vectors of 3D points from the 3D voxels.

It is clear that the local features extracted from voxels are more informative than that from 2D

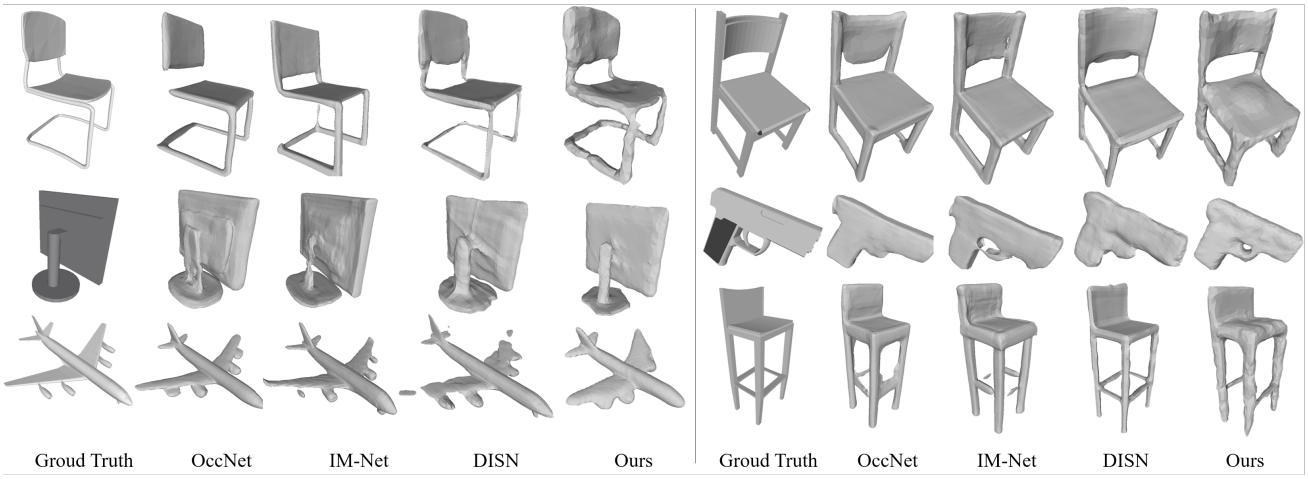

Figure 4. Visualization comparison between existing methods.

images. So the teacher network's feature vector can be used to supervise feature extraction of the SVR network.

Algorithm 1 shows how to compute loss from a single sample. In a training epoch, our method iterates over each sample in the dataset and uses Algorithm 1 to train the network. At the same time, back-propagation and parameter update are performed according to the loss. In Algorithm 1, we input image I, corresponding transformation matrix E, voxel of 3D model $mesh$ and polygonal mesh of 3D model $polygon$, then the algorithm will output loss, the sum of reconstruction loss and knowledge distillation loss, for this epoch. In the Algorithm 1, P is 3d point set, V is correspinding value set and P' is 2d point set mapped from P. Z_1, Z_2 are feature vectors, and o is implicit field prediction of reconstruction network.

Algorithm 1 Single Image Training Process

Input: Image I, corresponding transformation matrix E, voxel of 3D model $mesh$ and polygonal mesh of 3D model $polygon$

Output: Loss $loss$.

$P, V \leftarrow$ sample_points($mesh$)
$P' \leftarrow$ mapping(P, E)
$z_1 \leftarrow f_{CNN}(I, P'_i)$
$z_2 \leftarrow f_{3D-CNN}(mesh, P_i)$
$o \leftarrow f_{MLP}(z_1)$
$loss \leftarrow$ compute_loss(z_1, z_2, o, V_i)
return $loss$

4. Experiments

4.1. Citation dataset

Following the existing methods [13, 15, 24], we use 13 categories of ShapeNet-Core as our dataset. ShapeNet-Core is a densely annotated subset of ShapeNet covering 55 common object categories with about 51,300 unique 3D models. The input images are rendered by 3D-R2N2 [4] and their resolution is 137×137. The voxel dataset for the teacher network and data sampling is from HSP [25] and the resolution is $128 \times 128 \times 128$. To prepare the training data, we sample 10000 point-value pairs totally, of which 128 pairs are randomly sampled in 3D space and the others are sampled near the object surface. The sampling method is the same with IM-NET [24].

4.2. Implementation details

Network architecture. The implementation of U-Net [22] is the same with the origin one. The MLP, the reconstruction network, consists of five fully connected layers, using Leaky-ReLU [26] as activation function. For knowledge distillation, we use IF-Net [27] as our 3D teacher network. IF-Net takes sparse voxels as input, and extracts 5 local features in different scales for a 3D point. We concatenate all the local features as the target of the student network.

Training Details. We first train the SVR network only with reconstruction loss for 50 epochs, and then add the KD loss for another 30 epochs. The reconstruction loss is Mean Square Error (MSE) and the KD loss is L1 loss as commonly used. The Adam optimizer is used and the initial learning rate is 0.001. To generate surface from implicit field prediction, we apply Marching Cube algorithm [21] and the threshold is 0.5.

Metrics. Intersection of Union (IoU), Chamfer-L_1 Distance (CD) and Edge Chamfer Distance (ECD) [13] and DR-KFS [28] are used as quantitative metrics. To compute IoU metric, We voxelize reconstructed results and ground truths into $32 \times 32 \times 32$ voxels. To compute CD and ECD, We sample 4k and 16k points on the surface, respectively. DR-KFS is defined in [28] and the results are normalized into [0, 1].

4.3. Experiments results on ShapeNet-Core

Qualitative results. Figure 4 shows the qualitative comparison between occupancy network (OccMet) [13], IM-NET [24], DISN [15] and ours. All the methods can reconstruct the main body of 3D objects from a single image, but behave differently in details. One of the challenges is to reconstruct the thin connections in the 3D objects. OccNet and IM-Net both fail on the chairs and the screen. Especially in the first and last chairs, they can not generate complete surfaces. While our method can reconstruct most connections in these objects, but fails in the last chair. DISN perform well on the connection reconstruction, but it fails on the airplane. Overall, our method achieves comparable results with the existing methods, and even better performance on some details.

As we can see in Figure 4, our method reconstruct unsymmetrical results, such as the first chair and the airplane. This is a drawback of our method. Because we emphasize the local image feature in network design and training scheme, the network may be influenced by the perspective in some cases.

Table 1 lists the quantitative results. In addition to the baseline models of qualitative experiment, Pix2Vox++ [11] and AttSets [29] are added as comparison in quantitative experiment. ↑ represents

Table 1. Quantitative comparison between existing SVR methods.

	method	airplane	car	chair	display	lamp	rifle	table	mean
IOU(↑)	OccNET	**0.480**	0.570	0.358	0.439	0.254	0.427	0.461	0.461
	IM-NET	0.379	0.674	0.487	0.514	0.336	**0.468**	0.484	0.527
	DISN	0.328	0.672	0.301	0.358	0.189	0.197	0.105	0.360
	Pix2Vox++	0.413	0.630	0.435	0.324	**0.350**	0.417	0.305	0.411
	AttSets	0.398	0.612	0.403	0.301	0.334	0.433	0.306	0.398
	Ours	0.420	**0.686**	**0.505**	**0.527**	0.307	0.467	**0.509**	**0.545**
CD(↓)	OccNET	**0.461**	**0.368**	0.639	0.636	0.683	**0.414**	0.763	0.587
	IM-NET	0.574	0.650	0.919	0.907	0.802	0.556	0.979	0.797
	DISN	0.572	0.645	0.907	0.906	0.800	0.578	0.972	0.794
	Pix2Vox++	0.512	0.453	0.712	0.744	0.701	0.465	0.876	0.638
	AttSets	0.527	0.512	0.785	0.801	0.743	0.498	0.911	0.682
	Ours	0.581	0.394	**0.619**	**0.610**	**0.653**	0.419	**0.719**	**0.575**
ECD(↓)	OccNET	**0.423**	**0.288**	0.465	0.475	0.581	0.321	0.594	0.473
	IM-NET	0.522	0.370	0.627	0.641	0.695	0.478	0.750	0.589
	DISN	0.554	0.412	0.732	0.653	0.733	0.565	0.844	0.655
	Pix2Vox++	0.501	0.342	0.562	0.573	0.618	0.442	0.812	0.550
	AttSets	0.471	0.378	0.629	0.523	0.643	0.424	0.647	0.531
	Ours	0.493	0.329	**0.408**	**0.461**	**0.574**	**0.311**	**0.568**	**0.463**
DR-KFS(↓)	OccNET	**0.296**	**0.239**	0.325	0.366	0.402	0.291	0.438	0.337
	IM-NET	0.337	0.308	0.375	0.386	0.398	0.269	0.512	0.367
	DISN	0.324	0.313	0.392	0.403	0.422	0.401	0.524	0.397
	Pix2Vox++	0.311	0.297	0.355	0.372	0.407	0.335	0.503	0.369
	AttSets	0.324	0.325	0.373	0.401	0.398	0.283	0.497	0.372
	Ours	0.331	0.294	**0.323**	**0.341**	**0.390**	**0.257**	**0.429**	**0.316**

that the larger the metric is, the better, and ↓ is opposite. We display the results of 7 categories with the most shapes in ShapeNet-Core dataset. The OccNet shows great results on airplanes and cars. The most surfaces of these two categories are main body parts, where the OccNet does well. As we can find in Table 1, our method has best performance on most categories.

4.4. Ablation study on knowledge distillation

We conduct experiments to prove that the knowledge distillation loss is effectiveness. Table 2 compares the quantitative results of our methods without and with knowledge distillation. As we can see, KD loss can improve all the three metrics.

We also compare the effect of different architectures of teacher network. As described in 3.3, our 3D teacher network extracts feature vectors of different levels from the voxels input. We try to replace it with a simple one. The simple teacher is a naive 3D CNN also with 4 convolutional layers, but only outputs a single feature vector from the last layer. The results in Table 2 show that our teacher network is better than the simple one. The U-Net in our SVR pipeline leverage the multi-level feature extraction, so the teacher network should follow the same idea.

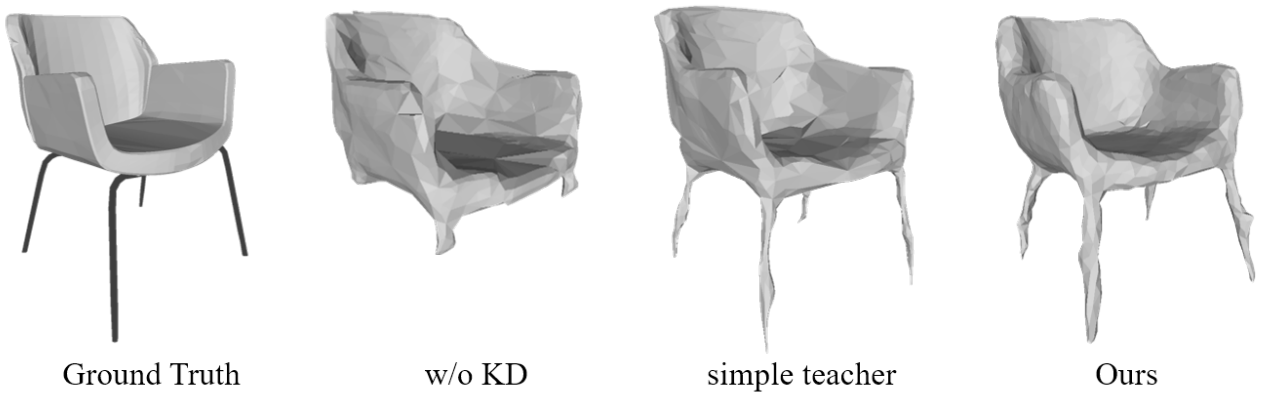

Figure 5. Qualitative comparison on 3D knowledge distillation.

Figure 5 shows a sample to explain the effect of knowledge distillation. Our model results are closest to the ground truth. In contrast, the network without knowledge distillation performs poorly in reconstructing details, and even some detailed parts, such as chair legs, are difficult to reconstruct. Since the simple teacher only outputs the feature vector of the last layer, although it can correctly identify and reconstruct the main part of the model, the texture features are tough. This is because it is easier for the shallow network to recognize the texture information in the input image, and the lack of feature vectors in the shallow network causes the simple teacher to reconstruct rough shapes. However, the multi-layer feature extraction in our method can solve this problem well.

Table 2. Quantitative comparison on 3D knowledge distillation.

	IOU(↑)	CD(↓)	ECD(↓)
w/o KD	0.539	0.595	0.471
simple teacher	0.541	0.579	0.464
Ours	0.545	0.575	0.463

4.5. The effect of the iterative process on the model performance

Our method is trained in an iterative manner. In order to study how the network can improve the performance in the iterative process, we show the change of the IoU of our model under different iteration rounds in the Figure 6, and give IoU values under some iteration rounds, as shown in the Table 3. As the iteration prohresses, the model performance gradually improves and becomes stable. The IoU value plummeted after the introduction of knowledge distillation. However, the performance of the network with knowledge distillation will gradually improve with iteration, and finally converge to a higher performance than the network without knowledge distillation.

4.6. Attempts to optimize skip connection

The skip connection is an important mechanism in U-Net [22]. It enriches the output feature map with texture information, which greatly benefits image segmentation task. However, SVR task needs more spatial information and these information can not be extracted by shallow convolutional layers. So we try to replace the skip connections with more complex modules, as shown in Figure 7. First,

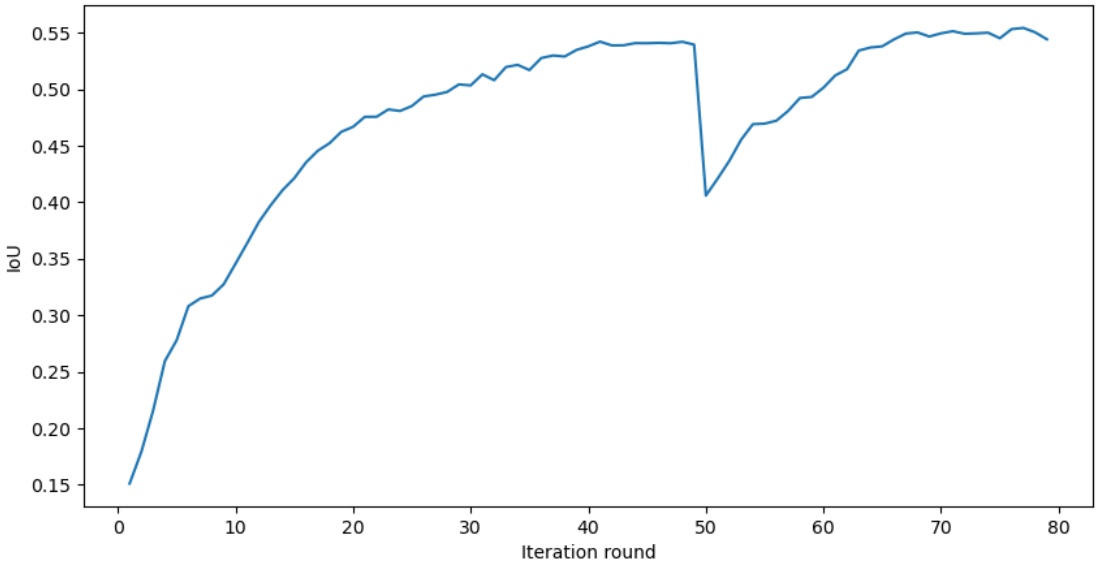

Figure 6. IoU in different iteration rounds.

Table 3. Statistics of IoU values under some iteration rounds. The first row represents the iteration round, and the second row represents the IoU value under this round. Due to the introduction of knowledge distillation in the 50th iteration, it can be seen that the IoU value has dropped significantly in this round of iterations. However, as the model continues to iterate, the IoU value of the network with knowledge distillation gradually rises and eventually surpasses the network without knowledge distillation.

iteration	1	10	20	30	40	49	50	60	70	80
IOU	0.115	0.345	0.466	0.503	0.537	0.539	0.405	0.501	0.549	0.544

we try to use a 5-layer CNN as the submodule. The simple CNN can increase the convolutional layers from the shallow layer to the deep one, and we expect more spatial information can be extracted.

We also try to apply Spatial Transform Network (STN) [30] in our network for more flexible feature extraction. STN can learn the spatial manipulations on the feature maps, so we expect it guide the network extract more discriminative features.

The results of attempts to optimize skip connection are listed in Table 4. Although 5-layer CNN and STN have some advantages, they do not outperform skip connection in our network. So we still use the skip connection submodule.

Table 4. Comparison between different submodule to replace skip connection.

	IOU(↑)	CD(↓)	ECD(↓)
5-layer CNN	0.520	0.621	0.593
STN	0.540	0.572	0.472
skip connection	0.545	0.575	0.463

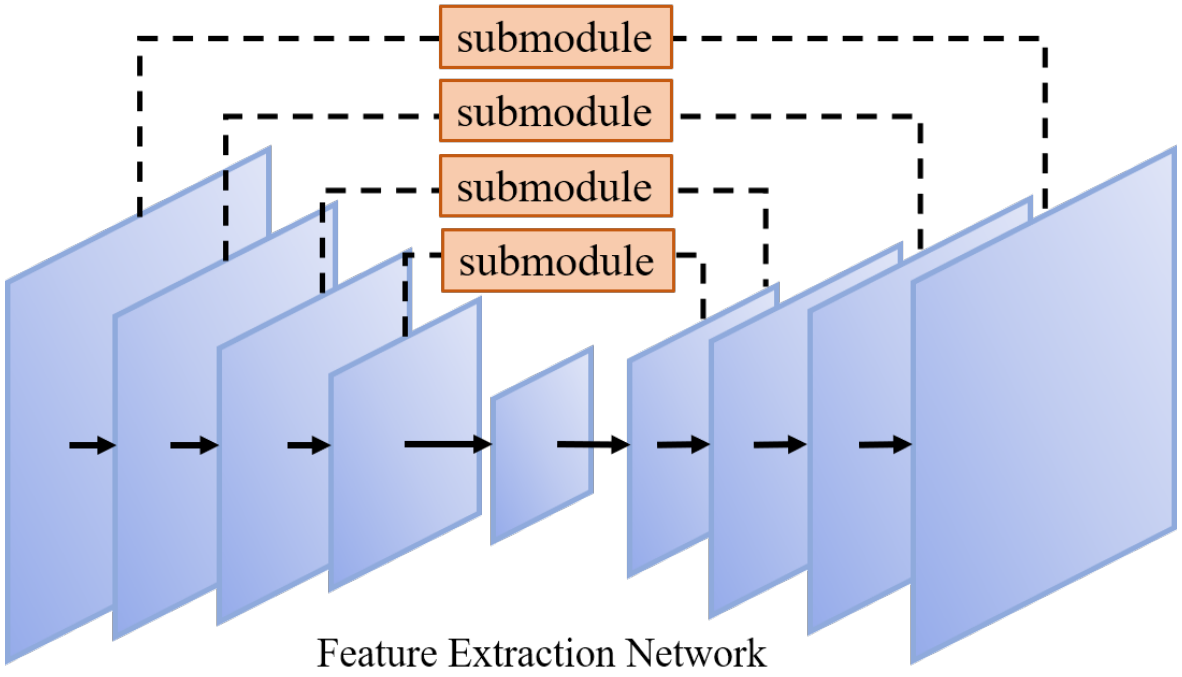

Figure 7. Feature extraction network structure replacing skip connections with submodules.

4.7. Time complexity

The training procedure is as follows. A 3D teacher network is trained first, followed by a 2D student network. Compared with the traditional method, our training time increases the time of training the teacher network, but we do not need to train the camera parameters prediction network, and the teacher network makes the training of the student network easier. Therefore, the overall time complexity will not increase much compared to the traditional method.

As for the test process, our model uses U-Net and MLP instead of the heavy network, so the test time will not be slower than the previous network. In a word, we pay more attention to the innovation of the mechanism than the performance of the network. However, our network structure is relatively lightweight, so the time complexity will not increase much compared to previous algorithms.

4.8. Experimental analysis

According to the analysis of the above tables and figures, we can draw the following conclusions:

1) From Figure 4 and the qualitative results, our method can accurately identify and reconstruct the overall shape of the model and reconstruct as many model details as possible. In contrast, OccNet and IM-Net often reconstruct parts that do not exist in the ground truth, or omit details of the ground truth. The DISN reconstruction results show poor continuity and also lack shape details. This indicates that the reconstruction results of our method are more in line with people's common sense and perform better in terms of vision.

2) From Table 1 and the quantitative results, our method outperforms the other three models under all four different metrics. Our method significantly outperforms OccNet and DISN on IoU and

ECD metrics, in addition, our method significantly outperforms IM-NET and DISN on CD metrics. Meanwhile, compared with the other three methods, our method still has advantages on DR-KFS metric.

3) Figure 5 and Table 2 show that the performance of the network without knowledge distillation becomes worse on all three metrics. Besides, the simple teacher using only the feature vector output from the last layer also performs worse than our method on all three metrics. This indicates that the introduction of knowledge distillation and multi-layer feature extraction is beneficial to the quality of the reconstructed results, demonstrating the effectiveness of both.

4) Figure 6 and Table 3 illustrate how the iterative process improves prediction results. It can be seen that with the increase of iteration rounds, the performance of the network on the IoU metric gradually improves and finally stabilizes around 0.539. Due to the introduction of knowledge distillation, the model performance plummeted at the 50th round. After that, as the training progresses, the performance of the network with knowledge distillation gradually improves, and finally stabilizes around 0.545, which is better than the network without knowledge distillation. This also proves the effectiveness of the introduction of knowledge distillation.

5) Due to the insufficiency of skip connection, we try to use 5-layer CNN and STN to replace the skip connection submodule. However, this attempt does not achieve the expected results, so our method still uses skip connection.

5. Discussion and conclusions

In this article, we propose IFKD model to simplify the mapping from 3D points to 2D pixels, so that local feature extraction can be very convenient and does not require additional camera information. Then, we perform knowledge distillation between a voxel-to-implicit field network and the SVR network to supervise the latter learning to extract a more informative feature vector. The experiments show that the proposed methods can improve the reconstruction performance.

To simplify the 3D-to-2D mapping, our method does not output aligned 3D objects, while the orientation is relative to the input image. This is a significant difference from the existing methods. However, we believe that the reconstruction quality is more important rather than the alignment.

Performing knowledge distillation between a 3D network and a SVR network is a novel design. There are still many things to explore. In the futre work, we will try more delicate knowledge distillation mechanism to help CNN extract 3D information from 2D image for SVR.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.61972014). The authors thank the anonymous referees and the editor for their valuable comments and suggestions.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. M. Li, H. Zhang, D2im-net: Learning detail disentangled implicit fields from single images, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 10246–10255. https://doi.org/10.1109/CVPR46437.2021.01011
2. S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, H. Li, Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2019), 2304–2314. https://doi.org/10.1109/ICCV.2019.00239
3. T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, M. Aubry, A papier-mâché approach to learning 3d surface generation, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2018), 216–224. https://doi.org/10.1109/CVPR.2018.00030
4. C. B. Choy, D. Xu, J. Gwak, K. Chen, S. Savarese, 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction, *Eur. Conf. Comput. Vision*, **9912** (2016), 628–644. https://doi.org/10.1007/978-3-319-46484-8_38
5. G. Yang, X. Huang, Z. Hao, M. Liu, S. Belongie, B. Hariharan, Pointflow: 3d point cloud generation with continuous normalizing flows, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2019), 4541–4550. https://doi.org/10.1109/ICCV.2019.00464
6. G. Chen, W. Choi, X. Yu, T. Han, M. Chandraker, Learning efficient object detection models with knowledge distillation, *Adv. Neural Inform. Proc. Syst.*, (2017), 30.
7. J. J. Park, P. Florence, J. Straub, R. Newcombe, S. Lovegrove, Deepsdf: Learning continuous signed distance functions for shape representation, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (2019), 165–174. https://doi.org/10.1109/CVPR.2019.00025
8. G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, preprint, arXiv: 1503.02531.
9. N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, Y. Jiang, Pixel2mesh: Generating 3d mesh models from single rgb images. in *Proceedings of the European conference on computer vision (ECCV)*, (2018), 52–67. https://doi.org/10.1007/978-3-030-01252-6_4
10. H. Xie, H. Yao, X. Sun, S. Zhou, S. Zhang, Pix2vox: Context-aware 3d reconstruction from single and multi-view images, preprint, arXiv: 1901.11153.
11. H. Xie, H. Yao, S. Zhang, S. Zhou, W. Sun, Pix2vox++: Multi-scale context-aware 3d object reconstruction from single and multiple images, *Int. J. Comput. Vision*, **128** (2020), 2919–2935. https://doi.org/10.1007/s11263-020-01347-6
12. L. P. Tchapmi, V. Kosaraju, H. Rezatofighi, I. Reid, S. Savarese, Topnet: Structural point cloud decoder, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019), 383–392. https://doi.org/10.1109/CVPR.2019.00047

13. L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, A. Geiger, Occupancy networks: Learning 3d reconstruction in function space. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, 4460–4470. https://doi.org/10.1109/CVPR.2019.00459

14. Z. Chen, A. Tagliasacchi, H. Zhang, Bsp-net: Generating compact meshes via binary space partitioning, preprint, arXiv:1911.06971.

15. Q. Xu, W. Wang, D. Ceylan, R. Mech, U. Neumann, Disn: Deep implicit surface network for high-quality single-view 3d reconstruction, *Adv. Neural Inform. Process. Syst.*, (2019), 32.

16. J. Bechtold, M. Tatarchenko, V. Fischer, T. Brox, Fostering generalization in single-view 3d reconstruction by learning a hierarchy of local and global shape priors, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 15880–15889. https://doi.org/10.1109/CVPR46437.2021.01562

17. Y. Duan, H. Zhu, H. Wang, L. Yi, R. Nevatia, L. J. Guibas, Curriculum deepsdf, *Eur. Conf. Comput. Vision*, (2020), 51–67. https://doi.org/10.1007/978-3-030-58598-3_4

18. A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, Y. Bengio, Fitnets: Hints for thin deep nets, preprint, arXiv: 1412.6550.

19. Q. Li, S. Jin, J. Yan, Mimicking very efficient network for object detection, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 7341–7349. https://doi.org/10.1109/CVPR.2017.776

20. T. Wang, L. Yuan, X. Zhang, J. Feng, Distilling object detectors with fine-grained feature imitation, preprint, arXiv: 1906.03609.

21. W. E. Lorensen, H. E. Cline, Marching cubes: A high resolution 3d surface construction algorithm, *ACM SIGGRAPH Comput. Graphics*, **21** (1987), 163–169. https://doi.org/10.1145/37402.37422

22. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, *Medical Image Computing and Computer-Assisted Intervention*, Springer, **9351** (2015), 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

23. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2009.

24. Z. Chen, H. Zhang, Learning implicit fields for generative shape modeling, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019), 5932-5941. https://doi.org/10.1109/CVPR.2019.00609

25. C. Häne, S. Tulsiani, J. Malik, Hierarchical surface prediction for 3d object reconstruction, preprint, arXiv: 1704.00710.

26. B. Xu, N. Wang, T. Chen, M. Li, Empirical evaluation of rectified activations in convolutional network, preprint, arXiv: 1505.00853.

27. J. Chibane, T. Alldieck, G. Pons-Moll, Implicit functions in feature space for 3d shape reconstruction and completion, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 6968–6979. https://doi.org/10.1109/CVPR42600.2020.00700

28. J. Jin, A. G. Patil, Z. Xiong, H. Zhang, Dr-kfs: A differentiable visual similarity metric for 3d shape reconstruction, *Eur. Conf. Comput. Vision*, (2020), 295–311. https://doi.org/10.1007/978-3-030-58589-1_18

29. B. Yang, S. Wang, A. Markham, N. Trigoni, Robust attentional aggregation of deep feature sets for multi-view 3d reconstruction, *Int. J. Comput. Vision*, **128** (2020), 53–73. https://doi.org/10.1007/s11263-019-01217-w

30. M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, Spatial transformer networks, in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, **2** (2015), 2017–2025. https://doi.org/10.5555/2969442.2969465

AIMS Press