Communications in Information and Systems
Volume 20, Number 4, 389–413, 2020

3D reconstruction using deep learning: a survey

Yiwei Jin, Diqiong Jiang, and Ming Cai*

Deep learning has remarkably improved the performance of many tasks in the computer vision community including 3D reconstruction. In this paper, we survey both classical and latest works of 3D reconstruction via deep learning. We divide all surveyed methods into three categories on the ground of the input modality: single RGB image based, multiple RGB images based and sketch based. Representations of output 3D shapes and specific goals of tasks are also taken into consideration in our classification. In addition, we overview datasets as well as evaluation metrics commonly used in current works. Finally, a discussion about potential directions of future research is provided.

1. Introduction

3D reconstruction is to represent 3D shapes of objects in given input. It is a fundamental challenge in wide applications ranging from remote sensing, navigation, 3D animation, medical assisting, and so on. Traditionally, mainstream methods of single image 3D reconstruction are based on certain assumptions of lighting and reflectance, thus are highly susceptible to the albedo, illumination, texture of the input. These methods capture shading, repetitive texture features from input images, while some others capture geometry information including contours, vertical and horizontal lines, vanishing points to reconstruct 3D shapes of the objects. And some use shooting lights to estimate the distance between viewpoint and the surface actively. However, these methods are very limited. With the development of deep learning techniques, both of the performance and the efficiency of 3D reconstruction have been remarkably improved. Early deep learning based methods take corresponding 3D groundtruths as supervisions, which are labor-intensive and difficult to get. Therefore, weaker supervisions projecting 3D information to 2D space are proposed to take place of 3D supervision. Including silhouette, depth maps, normal maps, etc., differentiable 2D supervisions are used to back-propagate the gradients from 2D loss functions

*Corresponding author.

in the networks. Furthermore, there are self-supervision methods minimizing the distance between 2D projections of reconstruction results and input images. And based on generative adversarial networks, some works are capable to reconstruct 3D shapes in an unsupervised manner. As for multi-view 3D reconstruction, photometric stereo and SFM (shape from motion) techniques are used to align the input images from different views previously. However, these methods are rather restricted in reconstructing objects in free views efficiently and accurately. Neural networks are first used to help alignment, but later they are used to reconstruct full shapes directly from input images. And the proposal of end-to-end structure has improved the efficiency of reconstruction frameworks. In addition, to deal with non-rigid object reconstructions, techniques of SFM, NRSFT (non-rigid shape from template) and prior models are employed in frameworks. In general, though 3D reconstruction technology has experienced a long development, the performance as well as the generalization capability has not reached saturation yet.

There are several crucial variations in the taxonomy of 3D reconstruction works, such as input modality, shape representation and network architecture. Basically, the input can be divided into two modalities: RGB image and sketch. Depth are considered as another modality in some literatures. However, depth based methods either take depth as an additional input or an intermediate of networks [1, 2, 3, 4], or focus on other tasks like segmentation and filling missing parts [5, 6, 7]. Therefore, in this paper, we only survey methods based on RGB image and sketch, without discussing depth input separately.

Due to its high availability in daily life, RGB image input has been deeply explored by the community. Among methods based on RGB input, some take a single image as the input of a network [8, 9, 10], while some others take a sequence of images [11, 12, 13], in forms of video frames and images of different viewpoints, as the input. Meanwhile, early sketch based methods usually take edge maps or standardized line-drawings as input, while thanks to the rise of end-to-end framework, hand-painted sketches that are more available to those non-professionals have been researched recently. To deal with the absence of information in sketch input, prior knowledge and generative adversarial networks are leveraged in these methods [14, 15, 16].

In addition to input, the representation of 3D shapes is of importance in reconstruction tasks. The representation impacts the architecture design as well as the performance of networks. Volumetric methods [17, 18, 19, 20, 2] represent 3D shapes by voxels in 3D grids. With a similarity of pixels in 2D images, volumetric networks can be easily extended from 2D convolutions.

However, it may lead to high memory consumption. Point cloud is a common representation of 3D shapes which is more memory saving. Point cloud based methods [21, 22] represent shapes by vertices in 3D coordinates. By implementing mappings onto the point clouds and dividing points into faces, meshes can be transformed from point clouds. Mesh based methods [23, 9] directly generalize meshes from input through mappings, despite suffering from topology and computation issues. Moreover, occupancy grids [24], octree [25, 26], parameters [27] and signed distance field (SDF) are also taken as representations of networks to reconstruct 3D shapes.

As for architecture, MLP (Multi-Layer Perceptron) is commonly used in initial methods based on neural networks. CNNs (Convolutional Neural Networks) are widely applied in reconstruction tasks for its appropriateness of dealing with 2D information. RNNs (Recurrent Neural Networks) are specifically employed for capturing sequential features in the input; while GANs (Generative Adversarial Networks) contribute to predicting missing information and improving the generalization ability of networks. Besides, recent works use GCNs (Graph Convolutional Networks) in some specific issues (e. g. face reconstructions) due to its suitability of dealing with non-Euclidean structure data.

In this paper, a comprehensive survey of 3D reconstruction using deep learning techniques is presented. We generally survey the reconstruction methods according to the input modality of networks, and organize each section under internal logic. The rest of this article is organized as follows. Section 2 includes common datasets and evaluation metrics used for 3D reconstruction. Section 3 surveys reconstruction methods based on single image input, while Section 4 focuses on multiple images. Section 5 reviews sketch based reconstructions. Finally, Section 6 discusses existing challenges of current works and potential directions of future researches.

2. Overview

2.1. Common dataset

This section reviews existing popular datasets used for 3D reconstruction.

ShapeNet [28] is a large-scale repository of shapes containing more than 50,000 CAD models organized in 55 classes. Annotations of semantic categories and attributes are also provided.

Pascal 3D+ [29] contains 12 classes of 3D rigid objects, with more than 3,000 objects per category. The dataset can be used for 3D object detection and pose estimation. Besides, it can be used as baselines for the community.

ObjectNet3D [30] consists of 100 categories and 90,127 images. Both 3D pose and shape annotations are provided for each 2D object in the images. It can also be used in proposal generation, 2D object detection and 3D pose estimation tasks.

KITTI [31] is a challenging dataset containing raw data captured by two video cameras and a laser scanner from rural areas and highways. It can be taken as real-world computer vision benchmarks in stereo, optical flow, visual odometry, 3D object detection and 3D tracking tasks.

Besides, there are some datasets used for specific reconstruction tasks, including BU-3DFE, Bosphorus, MICC, AFLW2000-3D for human face tasks and HumanEva [32] and Human3.6M [33] for human body tasks.

2.2. Metric

Common evaluation metrics used in 3D reconstruction are introduced as follows. Voxel IoU, MSE and cross-entropy loss are commonly used metrics in evaluating reconstruction performance of voxel representation.

MSE Mean Square Error (MSE) is the distance between a reconstructed 3D voxel shape and corresponding groundtruth.

$$MSE = \frac{\sum_{i=1}^{n} |p_i - g_i|^2}{n} \tag{1}$$

where $p_i \in [0, 1]$ denotes the predicted output at voxel i in a grid space, g_i is the corresponding groundtruth occupancy, and n is the total number of the voxels. Lower values indicate better reconstruction results.

Voxel IoU Intersection-over-Union (IoU) is commonly used in object detection tasks to measure the accuracy of a detecting algorithm. Extended to 3D reconstruction, Voxel IoU evaluates the accuracy of reconstructed shapes in voxel representation.

$$Voxel\ IoU = \frac{\sum_{i=1}^{n} I(p_i > t) I(g_i)}{\sum_{i=1}^{n} I(I(p_i > t)) + I(g_i))} \tag{2}$$

where $I(\cdot)$ is an indicator function and $t \in [0, 1]$ is a specified voxelization threshold. Higher values indicate better reconstruction results.

As for point cloud and mesh representation, we can evaluate the performance in several distance metrics between the reconstructed result and its groundtruth.

Average Euclidean Distance The average Euclidean distance is a basic metric in both 2D and 3D space. For point clouds, it is calculated by averaging the distance between each pair of corresponding points; while for meshes having faces information, the average Euclidean distance can be calculated along the normal as well.

Chamfer Distance Chamfer distance calculates the sum of distances between closest point pairs. The original asymmetric form is shown in equation (3).

$$CD(P, \hat{P}) = \frac{\sum_{p_i \in P} min_{\hat{p}_i \in \hat{P}} \|p_i - \hat{p}_i\|}{|P|} \tag{3}$$

where $P, \hat{P}$ are two point clouds. This formula can be extended into a symmetric form:

$$CD_{sym}(P, \hat{P}) = CD(P, \hat{P}) + CD(\hat{P}, P) \tag{4}$$

Besides, $\|\cdot\|$ can be changed to $\|\cdot\|_2$ (Euclidean distance) and other forms.

EMD Earth mover's distance (EMD) measures the distance between two probability distributions over a region in statistics. It is also used in image and singal processing tasks to quantitatively describe the difference.

$$EMD(P, \hat{P}) = min_{\phi: P \to \hat{P}} \sum_{pi \in P} \|p_i - \phi(p_i)\|_2 \tag{5}$$

where $\phi(\cdot)$ is a bijection from P to $\hat{P}$.

F-Score Given a threshold distance d, F-Score can be calculated from *precision* and *recall* of two point clouds. Particularly, *precision* is the percentage of reconstructed points lying within distance d to points on groundtruth surfaces. In reverse, *recall* is the percentage of groundtruth points lying within distance d to points on reconstructed surfaces.

$$F = \frac{\left(1 + \beta^2\right) Precision * Recall}{\beta^2 * Precision + Recall} \tag{6}$$

where we use β to balance the weights of *precision* and *recall*. When $\beta = 1$, the formula becomes

$$\frac{2}{F_1} = \frac{1}{Precision} + \frac{1}{Recall} \tag{7}$$

Table 1: Compare single image based methods over representation, target, supervision and architecture. GT: groundtruth

Ref	Representation	Target	Supervision	Architecture
[8]	voxel	object	2D GT, 3D GT	CNN
[10]	voxel	object	3D GT	GAN
[1]	voxel	novel object	3D GT	CNN
[20]	voxel	object with pose	bounding box, silhouette	GAN
[34]	voxel	human body	3D scan, prior model	CNN
[2]	point cloud	object	depth	CNN
[22]	point cloud	object	coarse points	CNN
[35]	point cloud	object	foreground mask	CNN
[9]	mesh	object	ellipsoid mesh	GCN
[36]	mesh	multiple human (non-rigid)	prior model	CNN
[37]	mesh	face caricature (non-rigid)	prior model	CNN

3. Single image reconstruction

In the past decades, single image based 3D reconstruction has developed from extracting geometry and texture features from limited types of images, to learning parameters of neural networks to estimate 3D shapes. Great improvements have been demonstrated in computational efficiency, reconstruction performance and generalization capability of 3D reconstruction.

The earliest deep learning based methods require real 3D shapes of target objects as supervision, which is very difficult to get at that time. Some researchers render images from CAD models to extend datasets, however problems occur that such synthesized data lead to lacks of generalization and authenticity in the reconstruction results. Some researchers take 2D and 2.5D projections of groundtruths as supervision and minimized reprojection losses during the learning process, such as contour, surface normal, etc. Later, methods that compare projections of the reconstructed results with the input to minimize the difference achieve to work with less supervision.

There are many factors influence both the procedure and the result of reconstruction. Table 1 shows the comparison between methods on 3D representation, reconstruction target, training supervision and network architecture. Despite having availability to 2D convolution, voxel representation shares a common challenge of high cost in memory, thus are limited in the scale of grids.

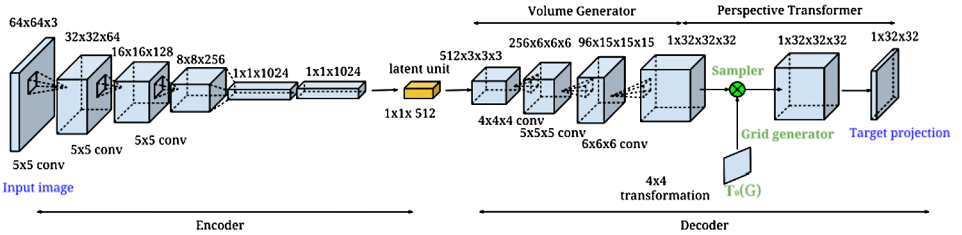

Figure 1: The architecture of Ref. [19].

3.1. Voxel representation

Voxel is one of the earliest 3D representations, which is very suitable for convolutional operations. Training their networks with 3D supervision, Zhang et al. [1] orderly predict a depth from given image under the same view and estimate a single-view spherical map from the depth. Then they use a voxel refinement network to integrate two projections and output a final reconstruction result. This work achieves generalizable and high-quality single image 3D reconstruction. Instead of requiring 3D groundtruth, some others using less supervision in the learning procedure. MarrNet [8] takes depth, normal map and silhouette as intermediate results to reconstruct 3D voxel shapes and use a reprojection consistency loss in the following procedure to estimate 3D shapes. Similarly, Yan et al. [19] propose a novel projection loss to learn 2D observation without 3D groundtruths. As shown in Figure 1, they use a 2D convolutional encoder, a 3D up-convolutional decoder and a perspective transformer network to reconstruct 3D voxels. They achieved state-of-the-art performance at that time. Different from the above works, Zhu et al. [20] reconstruct pose-aware 3D shapes from a single natural image. They refer to TL-embedding Network [38] and 3D-VAE-GAN [39] to construct the network, and minimize the reprojection loss over reprojected and ground truth silhouettes. Wu et al. [10] propose a framework that combines adversarial and volumetric convolutional networks to generate voxels from a probabilistic latent space in an unsupervised manner. They improve the generalization ability of the network, which is also studied in Refs. [21, 3, 10]. More recently, a lateset work TetraTSDF [34] proposes a tetrahedral volumetric representation of the human body and a method to retrieve detailed 3D body shapes wearing loose clothes from single 2D images. Based on the skinned multi-person linear model (SMPL) [40], they dilate, coarsen, up-sample and finally tetrahedralize the template model to

build a dense outer shell which is able to reconstruct details in lower resolutions compared with using standard rectangular voxels. To build the training data, pose and shape parameters of the outer shell are fitted to 3D scans and dense truncated signed distance fields (TSDF) are later produced by calculating TSDF values at the summits of each tetrahedra. While training, a novel part connection network of multiple feature layers is implemented to address the large memory consuming.

For multiple objects reconstruction [41, 42], CoReNet [43] jointly reconstructs multiple objects from a single image via a coherent reconstruction network. Going through a 2D encoder and a 3D decoder successively, all objects detected in the input image are represented in a single consistent 3D coordinate without intersection. And a ray-traced skip connection is introduced to ensure the physical accuracy. Specifically, they use a hybrid volume representation, in a maximum scale of grids of 128^3.

3.2. Point cloud representation

Point cloud is a sparse and memory saving representation compared with voxels. In early methods that take point clouds as the output of deep learning networks, Fan et al. [44] propose PointOutNet to reconstruct objects from single image. As shown in Figure 2, PointOutNet has a convolution encoder and two parallel predictor branches. The encoder takes in an image and a random vector that perturbs the prediction. And one of the branches is a fully-connected branch that captures complicated structures and another is a deconvolution branch that generates point coordinates. This network well exploits geometric continuity and is capable to generate smooth objects. Meanwhile, Lin et al. [2] utilize 2D convolutional operations to achieve higher efficiency. First, they use a generator to predict 3D structures at novel viewpoints from single image. They later use a pseudo-renderer to synthesize depth images of corresponding views, which are further used for joint 2D projection optimization. They predict denser point clouds of higher accuracy.

More recently, Zhang et al. [22] retrieve a nearest 3D shape as an extra input of the reconstruction network to generate fine-grained point clouds. They introduce an attention based 2D-3D fusion module into the network to integrate 2d and 3d features adaptively. Navaneet et al. [35] reconstruct point clouds from images of a certain category with corresponding foreground masks of each object. Based on an encoder-to-decoder structure, they use a geometric loss and a pose cycle consistency loss to train the networks in a self-supervised manner. They also achieve a multi-view and pose supervised approach in this work.

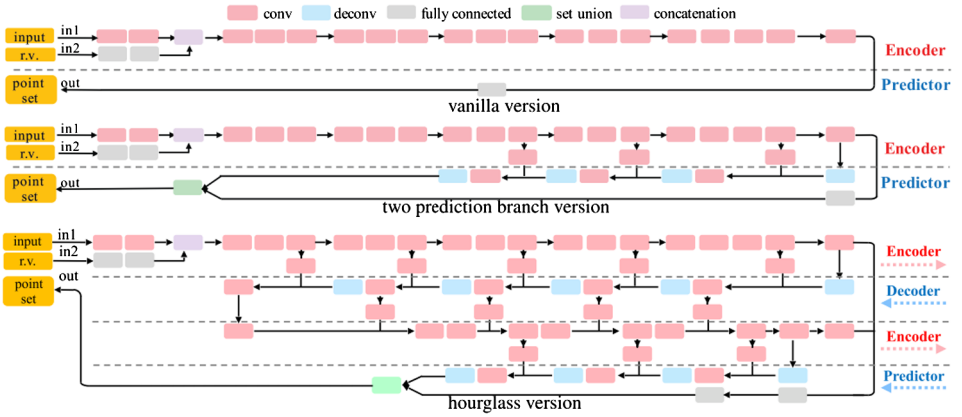

Figure 2: The architecture of Ref. [44].

3.3. Mesh representation

Previously, Kato et al. [45] reconstruct meshes from low-resolution images by employing an integrated mesh rendering network. They minimize the difference on 2D silhouettes between reconstructed objects and corresponding groundtruths. Pan et al. [23] propose ResMeshNet, a stacked framework consists of multiple MLP blocks, to reconstruct 3D meshes from single image. Specifically, they employ a shortcut connection between two blocks to retain the geometrical consistency. However, this work has difficulties in reconstructing smooth results with correct triangulation. To achieve higher fidelity, Pix2Mesh [9] uses a cascaded, graph-based convolutional network to reconstruct 3D meshes of rigid objects. The network captures perceptual features from the input image and deforms an ellipsoid progressively into the output geometry.

In addition to forementioned works, some others aim to reconstruct inherent deformations in non-rigid objects. Tasks of non-rigid reconstruction from single image usually require extra information of target objects, which can be either predicted during the process, or given as prior knowledge, such as skeleton structures and parameterized models.

Estimating poses and shapes of a human mesh from RGB images has extensive applications in daily life. Zeng et al. [36] propose a model-free method to establish dense correspondence between output mesh and local image features using a novel UV map. Jiang et al. [46] coherently reconstruct poses and shapes of multi-person in a single image based on zeng's work. Based on a prior model, Geo-PIFu [47] uses a deep implicit function to

represent clothed body shapes. It preserves global shape regularity as well as details of clothes by aligning and fusing local geometry and pixel features. And the forementioned work TetraTSDF [34] is able to reconstruct both poses and details of the human body dressed in loose clothes. On the other hand, Zhang et al. [37] address a quite different issue. Given a caricature, they detect 2D landmarks on the faces and use them to generate exaggerate, distorted face shapes. They are capable to reconstruct face shapes from caricatures through a nonlinear parametric model.

3.4. Other representation

Other than the representations mentioned above, 2.5D [3, 48] representation as well as occupancy grid [25, 26] and implicit surface [49, 50] based volumetric representation are also utilized to represent reconstructed 3D shapes.

Depth Liu ei al. [48] generate depth maps of input images through recurrent attentional networks and recognize interested objects from the depths using CNN. Zhou et al. [3] utilize mirror planes to predict depth maps of self-symmetry objects by finding corresponding pixel pairs in the image. They introduce cost volumes into the reconstruction procedure to preserving the geometry and simultaneously generate confidence vector and depth estimation. Wu et al. [51] also leverage symmetry in their work and achieve unsupervised single image reconstruction. Assuming that the object in an input image is horizontally symmetric, they use an autoencoder to estimate the light direction, symmetric depth maps and albedos of a frontal view and predict a canonical like of the input object through these results, followed by transforming the canonical image into the actual view to evaluate the difference.

Implicit Surface OccNet [49] implicitly represents 3D shapes as the continuous decision boundary of a deep network classifier and achieves relatively higher resolution and lower memory occupation. UCLID-Net [50] proposes a multi-layer network architecture to extract geometry features and represents 3D shapes in an Euclidean preserving latent space. Michalkiewicz et al. [52] use a CNN based decoder to predict SDF representations from a latent space. Later, it is demonstrated that adopting principal component analysis on SDF in reconstruction process allows higher resolution and quality [53]. SPSG [54] leverages 2D view-guided synthesis in reconstructing 3D objects of TSDF from incomplete single RGB-D scan. This network is capable to

reconstruct geometry and color respectively in a self-supervised way. Similar with SPSG in reconstructing both shapes and colors, Im2Avatar [55] represents 3D shapes in occupancy grids and colors in volumes. Recently, TSDF based representations (e.g., combined with volume or octree [34, 56]) have shown an advantage in dealing with interference (e.g., overlap and noise) and modeling continuous surfaces efficiently.

In general, reconstruction 3D objects from single image is challenging due to inherent ambiguity and self-occlusion in single view input. Although both performance and efficiency have been improved greatly over the past few years, more research is needed in reconstructions from single image.

4. Multiple image reconstruction

When images of different views are taken as input of networks, the inherent ambiguity of the object keeps reducing and obscured parts are supplemented.

Traditionally, there are two main types of multi-view reconstruction. One is to reconstruct a still object from images of two or more views, while the other is to reconstruct the 3D shape of a moving object from a video or multiple frames. Both of these methods estimate camera pose and corresponding shape from an image and align the partial 3D shapes into a full one. Therefore, the difficulty lies in pose estimating and 3D alignment. Deep learning techniques are first introduced into multi-image reconstruction to address this issue. Later, deep neural networks are used to directly generate 3D shapes from input images. And the employment of end-to-end structure greatly reduce the time consume of the reconstruction process. Table 2 shows the comparison between multiple image based reconstruction methods in input modality, 3D representation and reconstruction target.

4.1. Rigid reconstruction

Previously, Xie et al. [17] propose an encoder-decoder structure framework Pix2Vox++ based on RNNs to generate a coarse volume for each input image. As shown in Figure 3, They fuse all the coarse volumes through a multi-scale context-aware fusion module and adopt a refiner to correct the fused volume. Inspired by the standard LSTM framework, 3D-R2N2 [24] outputs 3D shapes in occupancy grids with the only supervision of bounding box. It unifies single and multi-view reconstruction in an encoder-LSTM-decoder structure. The 3D convolutional LSTM updates hidden representations selectively through input gates and forget gates. It effectively handles the self-occlusion and incrementally refines the reconstruction result as more

Table 2: Compare multi-image based methods over representation, input and target

Ref	Representation	Input	Target
[17]	voxel	image(s)	object
[24]	voxel	image(s)	object
[57]	voxel	images	object
[11]	voxel	images	object
[58]	point cloud	image(s)	novel object
[12]	point cloud	images	object
[27]	mesh	image(s)	face parameters
[59]	mesh	images	object (pose) hand (pose, shape)
[13]	SDF	images	objects
[54]	TSDF	RGBD scan	objects
[56]	TSDF	depth maps	object

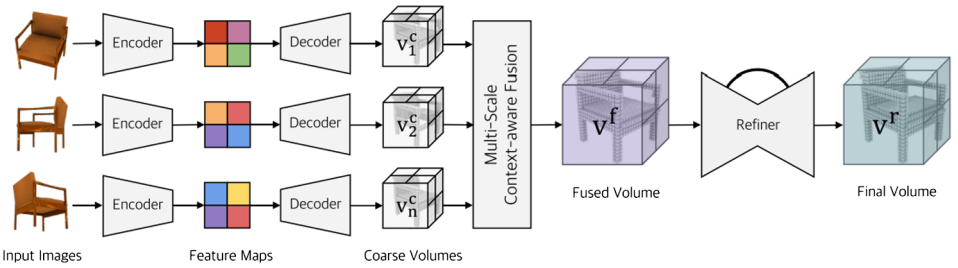

Figure 3: The architecture of Ref. [17].

observations being taken in. However, despite capability to retain previous observations, methods based on such structure may fail when given similar images and are limited to retain features in early inputs.

Given a RGB video, FroDO [13] use a coarse-to-fine framework to reconstruct multiple 3D meshes with sparse point clouds and precomputed camera poses. The network consists a CNN based encoder to predict a 64D parameter for each different object detected in the image sequence, as well as two decoders to predict object shapes as points and SDF. The final reconstructed result is incrementally generated through refining and fusion processes. Cascaded coarse-to-fine structure is a reasonable choice to reconstruct high-resolution geometries with low memory consuming. Cao et al. [56] use cascaded, multi-stage networks to infer missing surface areas and refine geometric details from incomplete and noisy depth maps. They first integrate raw depth scans into a TSDF volume, followed by implementing two 3D fully convolutional networks to respectively regress a complete

low-resolution TSDF and infer detailed high-resolution patches to refine the regressed TSDF. Note that they substitute modified networks in OctNet [25] for convolution and pooling layers in UNet [60] to store TSDF volumes in the structure of octree.

To address alignment without 3D supervisions, Banani et al. [11] propose learning two networks: one is to estimate the 3D shape of an object from two images of different viewpoints with corresponding pose vectors, and predict the object's appearance from a third view; and the other evaluates the misalignment of two views. While testing, they predict a transformation that best aligns the bottleneck features of two input images. Their networks also work on generalizing unseen objects.

Moreover, there are some specific tasks focusing on reconstructing buildings [61, 62, 63], large-scale scenes [64, 65, 66] and reconstructing under dynamic views [67, 68, 69]. Researchers use scanners, depth cameras and other devices to generate point clouds and depth maps separately over frames. Difficulties usually lie in distance standardizing and camera pose estimation [70, 71].

4.2. Non-rigid reconstruction

As for non-rigid reconstruction, Zuo et al. [72] utilize the SMPL model to build up a human avatar from a sparse RGBD video. They perform an initial pairwise alignment over every adjacent two frames of the video and generate a full 3D shape through a global non-rigid registration procedure. Besides, they also present a texture mapping method to texture the reconstructed human meshes by deforming textures captured from several frames. Differently, Ref. [72, 73] aim to track the motion of human and dynamic objects.

It is noticeable that tasks of hand-object image reconstruction can be quite different. For one hand, when the hand is interacting with the object (like grab and hold), there is probably large occlusion in the images. Works [74, 75] focus on interaction tasks and are able to reconstruction shapes and poses of hands. For the other hand, the poses of hand and object are intrinsically relevant. Some researchers [76, 77, 78] benefit from training to learn poses of hands and objects jointly. And a recent work [59] introduces a novel photometric loss into a single feed-forward neural network to estimate poses and shapes of both hands and objects jointly.

Furthermore, face reconstruction [79, 80] is common filed in non-rigid 3D reconstruction as well. Compared with MultiViewFace [27] that reconstructs a face mesh from multi-view images based on RNN, DeepFaceFlow

[79] predicts 3D flows of human faces using convolutional networks. They reconstruct dense high-quality dynamic face shapes with 3d groundtruth scans from pairs of frames in monocular facial videos. And for real-time reconstructing, Thomas [80] proposes a detailed 3D face animating system based on a RGBD camera. They implement rigid reconstruction on the first neutral face frame to reconstruct face shapes in the blendshape representation and track the deformation of shape and texture in the following frames. They overcome the lack of geometry details by augmenting blendshape meshes with a pair of deviation and color images.

Generally, among rigid object and scene reconstruction, current works mainly focus on improving the resolution of reconstruction results and enhancing network generalization capability. As for the reconstruction of non-rigid objects, due to their internal structure variability, prior knowledge are employed to estimate deformation of templates and parameters of models. Consequently, such methods are capable to reconstruct 3D representations of average shape, but are limited in reconstructing details and characteristics.

5. Sketch reconstruction

Sketch is another type of input in 3D reconstructions, which contains least information for human to perceive full 3D shapes, such as edge map, binary silhouette, line drawing and so on. 3D reconstruction based on sketch input has been studied in past several years. Due to lacks of visual information, previous works usually require additional supervision, including surface normal and multiple viewpoints, during the learning procedure.

Delanoy et al. [15] learn 3D voxels from edge maps of multiple views. The network first predicts an initial reconstruction via a single-view CNN and refines the output with the increment of edge maps from different viewpoints. Xin et al. [81] predict editable 3D shapes of cuboids and cylinders from generated instance masks with semantic labels. They uses an instance-aware segmentation network and a deformable convolutional network to predict labeled part-level masks from RGB images. Similarly with Ref. [45], their editing is implemented on 2D supervision by adjusting the extracted trajectory axis on these masks.

However, these works take only professional, accurate sketches as input. To adjust the network for input with more geometrical distortion, Wang et al. [16] propose a framework to reconstruct 3D volumes with poses from single freehand sketch. They employ a generative model to synthesize sketches from object images as training data. And a standardization module is adopted to transform sketch styles and enhance the generalization ability

of the network. Furthermore, some works are proposed to reduce the supervision. PrGANs [14] learns a probabilistic distribution over 3D shapes from a collection of 2D binary sketches of a certain category in an unsupervised manner. Based on GANs, it consists of a generator to generate 3D voxels, a projection module to render2D binary sketches from certain view, and a discriminator to classify whether a binary image is real. With a similar architecture of 3D-GAN [10], PrGANs achieves better reconstruction performance but fails to capture concave interior structures.

Notablely, human related freehand sketch is getting its popularity during recent years in both 2D [82, 83, 84, 85, 86] and 3D reconstruction [87, 88, 89, 90]. Early in 2009, Sketch2Photo [82] was proposed to generate photo-realistic pictures from single sketch image. Han et al. [87] propose a CNN based modeling system to reconstruct 3D meshes from single freehand caricature through a bilinear model of shape and expression variations. The framework also supports gesture based user interactions to further manipulate the face models. Later, the framework is further extended [89] to exaggerate faces in 3D meshes. Given a face photo, the new framework is able to reconstruct a 3D face mesh and edit it to approximate a freehand caricature.

In this section, we discuss about the 3D reconstruction methods based on undistorted and freehand sketches. Targets of sketch based reconstructions are generally objects with strong geometric features. Techniques of variational auto-encoders and generative networks are commonly employed to address the issue of cue deficiency. Noticeably, freehand sketch, especially sketch of human faces, has gained popularity during the past several years. These methods usually rely on prior models and generative adversarial networks to obtain plausible results. Generally, reconstructing credible shapes from freehand sketch needs further research in the future.

6. Conclusions

In this paper, we provide a comprehensive survey of deep learning based 3D reconstruction. We review extensive methods in the last decade and categorize them according to different input modalities: single image, multiple images and sketch input. Each category is organized under well-designed structure. In addition to the taxonomy, advantages and limitations are also discussed in each category. Moreover, we focus on latest works to follow the current progress and mainstream of the community.

The representation of reconstructed shapes also infects the architecture of networks. Voxel representation suffers a limitation of low resolution due

to its high memory consume; and point cloud representation lacks details and characteristics especially in tasks requiring prior models. Furthermore, in order to handle with the lack of training data, some researchers adopt methods of 2D and weaker supervision to avoid using 3D groundtruth; while to reconstruct objects from unseen categories, some researchers adjust the architectures to enhance the generalization ability of the networks. More researches are needed to further study the forementioned issues.

Meanwhile, in order to make reconstruction tasks more accessible to non-professionals, sketch input has been increasingly popular. We consider it a potential direction of future researches that to express the input with less constrains in wider range of reconstruction tasks. In addition, there are many special tasks in 3D reconstruction researches. Reconstructing shapes with specific attributes is in demand, such as the identity and expression attributions of human face, pose of human body, gesture in interaction tasks, focus in medical field, etc. We therefore believe that approaches of more universality and investigations on specific tasks are required in the future.

Acknowledgement

The research is supported in part by National Natural Science Foundation of China (NSFC) (61972342) and the Science and Technology Department of Zhejiang Province (2018C01080).

References

[1] X. Zhang, Z. Zhang, C. Zhang, J. Tenenbaum, B. Freeman, and J. Wu, "Learning to reconstruct shapes from unseen classes," in *Advances in Neural Information Processing Systems*, 2018, pp. 2257–2268.

[2] C.-H. Lin, C. Kong, and S. Lucey, "Learning efficient point clou generation for dense 3d object reconstruction," in *Thirty-Second AAA Conference on Artificial Intelligence*, 2018.

[3] Y. Zhou, S. Liu, and Y. Ma, "Learning to detect 3d reflection symme try for single-view reconstruction," *arXiv preprint arXiv:2006.10042*, 2020.

[4] A. Thai, S. Stojanov, V. Upadhya, and J. M. Rehg, "3d reconstruction of novel object shapes from single images," *arXiv preprint* arXiv:2006.2020, 2020.

[5] C. Zou, E. Yumer, J. Yang, D. Ceylan, and D. Hoiem, "3d-prnn: Gen erating shape primitives with recurrent neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 900–909.

[6] B. Yang, H. Wen, S. Wang, R. Clark, A. Markham, and N. Trigoni, "3d object reconstruction from a single depth view with adversarial learning," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 679–688.

[7] B. Yang, S. Rosa, A. Markham, N. Trigoni, and H. Wen, "Dense 3d object reconstruction from a single depth view," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 12, pp. 2820–2834, 2018.

[8] J. Wu, Y. Wang, T. Xue, X. Sun, B. Freeman, and J. Tenenbaum, "Marrnet: 3d shape reconstruction via 2.5d sketches," in *Advances in Neural Information Processing Systems*, 2017, pp. 540–550.

[9] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang, "Pixel2mesh: Generating 3d mesh models from single rgb images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 52–67.

[10] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum, "Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling," in *Advances in Neural Information Processing Systems*, 2016, pp. 82–90.

[11] M. E. Banani, J. J. Corso, and D. F. Fouhey, "Novel object viewpoint estimation through reconstruction alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3113–3122.

[12] D. J. Rezende, S. A. Eslami, S. Mohamed, P. Battaglia, M. Jaderberg, and N. Heess, "Unsupervised learning of 3d structure from images," in *Advances in Neural Information Processing Systems*, 2016, pp. 4996–5004.

[13] K. Li, M. Riinz, M. Tang, L. Ma, C. Kong, T. Schmidt, I. Reid, L. Agapito, J. Straub, S. Lovegrove, et al., "Frodo: From detections to 3d objects," *arXiv preprint* arXiv:2005.05125, 2020.

[14] M. Gadelha, S. Maji, and R. Wang, "3d shape induction from 2d views of multiple objects," in *2017 International Conference on 3D Vision (3DV)*, IEEE, 2017, pp. 402–411.

[15] J. Delanoy, M. Aubry, P. Isola, A. A. Efros, and A. Bousseau, "3d sketching using multi-view deep volumetric prediction," in *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, vol. 1, no. 1, pp. 1–22, 2018.

[16] J. Wang, J. Lin, Q. Yu, R. Liu, Y. Chen, and S. X. Yu, "3d shape reconstruction from free-hand sketches," *arXiv preprint* arXiv:2006.09694, 2020.

[17] H. Xie, H. Yao, S. Zhang, S. Zhou, and W. Sun, "Pix2vox++: Multi-scale context-aware 3d object reconstruction from single and multiple images," *International Journal of Computer Vision*, pp. 1–17, 2020. MR4156269

[18] A. Johnston, R. Garg, G. Carneiro, I. Reid, and A. van den Hengel, "Scaling cnns for high resolution volumetric reconstruction from a single image," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 939–948.

[19] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee, "Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision," in *Advances in Neural Information Processing Systems*, vol. 1, pp. 1696–1704.

[20] R. Zhu, H. Kiani Galoogahi, C. Wang, and S. Lucey, "Rethinkin reprojection: Closing the loop for pose-aware shape reconstruction from a single image," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 57–65.

[21] T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry "Atlasnet: A papier-mache approach to learning 3d surface generation," in *Proceedings of the IEEE Conference on Computer Vision an Pattern Recognition*, 2018, pp. 216–224.

[22] Y. Zhang, Z. Liu, T. Liu, B. Peng, and X. Li, "Realpoint3d: An efficien generation network for 3d object reconstruction from a single image, *IEEE Access*, vol. 7, pp. 57539–57549, 2019.

[23] J. Pan, J. Li, X. Han, and K. Jia, "Residual meshnet: Learning to deform meshes for single-view 3d reconstruction," in *2018 International Conference on 3D Vision (3DV)*, IEEE, 2018, pp. 719–727.

[24] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3d-r2n2: A unified approach for single and multi-view 3d object reconstruction, in *European Conference on Computer Vision*, Springer, 2016, pp. 628–644.

[25] G. Riegler, A. Osman Ulusoy, and A. Geiger, "Octnet: Learning deep 3d representations at high resolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3577–3586.

[26] M. Tatarchenko, A. Dosovitskiy, and T. Brox, "Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2088–2096.

[27] P. Dou and I. A. Kakadiaris, "Multi-view 3d face reconstruction with deep recurrent neural networks," *Image and Vision Computing*, vol. 80, pp. 80–91, 2018.

[28] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al., "Shapenet: An information-rich 3d model repository," *arXiv preprint* arXiv:1512.03012, 2015.

[29] Y. Xiang, R. Mottaghi, and S. Savarese, "Beyond pascal: A benchmark for 3d object detection in the wild," in *IEEE Winter Conference on Applications of Computer Vision*, IEEE, 2014, pp. 75–82.

[30] Y. Xiang, W. Kim, W. Chen, J. Ji, C. Choy, H. Su, R. Mottaghi, L. Guibas, and S. Savarese, "Objectnet3d: A large scale database for 3d object recognition," in *European Conference on Computer Vision*, Springer, 2016, pp. 160–176.

[31] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2012, pp. 3354–3361.

[32] L. Sigal, A. O. Balan, and M. J. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *International Journal of Computer Vision*, vol. 87, no. 1–2, p. 4, 2010.

[33] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.

[34] H. Onizuka, Z. Hayirci, D. Thomas, A. Sugimoto, H. Uchiyama, and R.-i. Taniguchi, "Tetratsdf: 3d human reconstruction from a single image with a tetrahedral outer shell," in *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, 2020, pp. 6011–6020.

[35] K. Navaneet, A. Mathew, S. Kashyap, W.-C. Hung, V. Jampani, and R. V. Babu, "From image collections to point clouds with self-supervised shape and pose networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1132–1140.

[36] W. Zeng, W. Ouyang, P. Luo, W. Liu, and X. Wang, "3d human mesh regression with dense correspondence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7054–7063.

[37] J. Zhang, H. Cai, Y. Guo, and Z. Peng, "Landmark detection and 3d face reconstruction for caricature using a nonlinear parametric model," *arXiv preprint* arXiv:2004.09190, 2020.

[38] R. Girdhar, D. F. Fouhey, M. Rodriguez, and A. Gupta, "Learning predictable and generative vector representation for objects," in *European Conference on Computer Vision*, Springer, 2016, pp. 484–499.

[39] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum, "Learning a probabilistic latent space of object shapes via 3d generative adversarial modeling," in *Advances in Neural Information Processing Systems*, 2016, pp. 82–90.

[40] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black "Smpl: A skinned multi-person linear model," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, pp. 1–16, 2015.

[41] A. Kundu, Y. Li, and J. M. Rehg, "3d-rcnn: Instance-level 3d object reconstruction via render-and-compare," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3559–3568.

[42] H. Izadinia, Q. Shan, and S. M. Seitz, "Im2cad," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5134–5143.

[43] S. Popov, P. Bauszat, and V. Ferrari, "Corenet: Coherent 3d scen reconstruction from a single rgb image," *arXiv preprint* arXiv:2004.12989, 2020.

[44] H. Fan, H. Su, and L. J. Guibas, "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, 2017, pp. 605–613.

[45] H. Kato, Y. Ushiku, and T. Harada, "Neural 3d mesh renderer," in *Proceedings of the IEEE Conference on Computer Vision and Patter Recognition*, 2018, pp. 3907–3916.

[46] W. Jiang, N. Kolotouros, G. Pavlakos, X. Zhou, and K. Daniilidk, "Coherent reconstruction of multiple humans from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5579–5588.

[47] T. He, J. Collomosse, H. Jin, and S. Soatto, "Geo-pifu: Geometry an pixel aligned implicit functions for single-view human reconstruction, *arXiv preprint* arXiv:2006.08072, 2020.

[48] M. Liu, Y. Shi, L. Zheng, K. Xu, H. Huang, and D. Manocha, "Recurrent 3d attentional networks for end-to-end active object recognition," *Computational Visual Media*, vol. 5, no. 1, pp. 91–104, 2019.

[49] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4460–4470.

[50] B. Guillard, E. Remelli, and P. Fua, "Uclid-net: Single view reconstruction in object space," *arXiv preprint* arXiv:2006.03817, 2020.

[51] S. Wu, C. Rupprecht, and A. Vedaldi, "Unsupervised learning of probably symmetric deformable 3d objects from images in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1–10.

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

[53] M. Michalkiewicz, E. Belilovsky, M. Baktashmotlagh, and A. Eriksson, "A simple and scalable shape representation for 3d reconstruction," *arXiv preprint* arXiv:2005.04623, 2020.

[54] A. Dai, Y. Siddiqui, J. Thies, J. Valentin, and M. Nießner, "Spsg: Self-supervised photometric scene generation from rgb-d scans," *arXiv preprint* arXiv:2006.14660, 2020.

[55] Y. Sun, Z. Liu, Y. Wang, and S. E. Sarma, "Im2avatar: Colorful 3d reconstruction from a single image," *arXiv preprint* arXiv:1804.06375, 2018.

[56] Y.-P. Cao, Z.-N. Liu, Z.-F. Kuang, L. Kobbelt, and S.-M. Hu, "Learning to reconstruct high-quality 3d shapes with cascaded fully convolutional networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 616–633.

[57] J. Gwak, C. B. Choy, M. Chandraker, A. Garg, and S. Savarese, "Weakly supervised 3d reconstruction with adversarial constraint," in *2017 International Conference on 3D Vision (3DV)*, IEEE, 2017, pp. 263–272.

[58] M. A. Bautista, W. Talbott, S. Zhai, N. Srivastava, and J. M. Susskind, "On the generalization of learning-based 3d reconstruction," *arXiv preprint* arXiv:2006.15427, 2020.

[59] Y. Hasson, B. Tekin, F. Bogo, I. Laptev, M. Pollefeys, and C. Schmid, "Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 571–580.

[60] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2015, pp. 234–241.

[61] B. Xiong, M. Jancosek, S. O. Elberink, and G. Vosselman, "Flexible building primitives for 3d building modeling," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 101, pp. 275–290, 2015.

[62] B. Xu, W. Jiang, and L. Li, "Hrtt: A hierarchical roof topology structure for robust building roof reconstruction from point clouds," *Remot Sensing*, vol. 9, no. 4, p. 354, 2017.

[63] B. Xu, X. Zhang, Z. Li, M. Leotta, S.-F. Chang, and J. Shan, "Deep learning guided building reconstruction from satellite imagery-derived point clouds," *arXiv preprint* arXiv:2005.09223, 2020.

[64] J. Zhang, L. Tai, J. Boedecker, W. Burgard, and M. Liu, "Neural slam: Learning to explore with external memory," *arXiv preprint* arXiv:1706.09520, 2017.

[65] P. Mirowski, R. Pascanu, F. Viola, H. Soyer, A. J. Ballard, A. Baninc M. Denil, R. Goroshin, L. Sifre, K. Kavukcuoglu, et al., "Learning to

navigate in complex environments," *arXiv preprint* arXiv:1611.03673, 2016.

[66] S. Yang, B. Li, Y.-P. Cao, H. Fu, Y.-k. Lai, L. Kobbelt, and S.-m Hu, "Noise-resilient reconstruction of panoramas and 3d scenes using robot-mounted unsynchronized commodity rgb-d cameras," *ACM Transactions on Graphics*, 2020.

[67] J. Schwartz, H. Zheng, M. Hanwell, Y. Jiang, and R. Hovden, "Dynamic compressed sensing for real-time tomographic reconstruction," *arXiv preprint* arXiv:2005.01662, 2020.

[68] S. Meerits, D. Thomas, V. Nozick, and H. Saito, "Fusionmls: Highly dynamic 3d reconstruction with consumer-grade rgb-d cameras," *Computational Visual Media*, vol. 4, no. 4, pp. 287–303, 2018.

[69] P. Wang, L. Liu, N. Chen, H.-K. Chu, C. Theobalt, and W. Wang, "Vid2curve: Simultaneously camera motion estimation and thin structure reconstruction from an rgb video," *arXiv preprint* arXiv:2005.03372, 2020.

[70] Y. Nakajima and H. Saito, "Robust camera pose estimation by viewpoint classification using deep learning," *Computational Visual Media*, vol. 3, no. 2, pp. 189–198, 2017.

[71] C. Wang and X. Guo, "Feature-based rgb-d camera pose optimization for real-time 3d reconstruction," *Computational Visual Media*, vol. 3, no. 2, pp. 95–106, 2017.

[72] X. Zuo, S. Wang, J. Zheng, W. Yu, M. Gong, R. Yang, and L. Cheng, "Sparsefusion: Dynamic human avatar modeling from sparse rgbd images," *arXiv preprint* arXiv:2006.03630, 2020.

[73] M. Shi, K. Aberman, A. Aristidou, T. Komura, D. Lischinski, D. Cohen-Or, and B. Chen, "Motionet: 3d human motion reconstruction from monocular video with skeleton consistency," *arXiv preprint* arXiv:2006.12075, 2020.

[74] D. Tzionas, L. Ballan, A. Srikantha, P. Aponte, M. Pollefeys, and J. Gall, "Capturing hands in action using discriminative salient points and physics simulation," *International Journal of Computer Vision*, vol. 118, no. 2, pp. 172–193, 2016. MR3508214

[75] H. Joo, T. Simon, and Y. Sheikh, "Total capture: A 3d deformation model for tracking faces, hands, and bodies," in *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition, 2018, pp. 8320–8329.

[76] M. Kokic, D. Kragic, and J. Bohg, "Learning to estimate pose and shape of hand-held objects from rgb images," *arXiv preprint* arXiv:1903.03340, 2019.

[77] S. Sridhar, F. Mueller, M. Zollhofer, D. Casas, A. Oulasvirta, and C. Theobalt, "Real-time joint tracking of a hand manipulating an object from rgb-d input," in *European Conference on Computer Vision*, Springer, 2016, pp. 294–310.

[78] Y. Chen, Z. Tu, D. Kang, R. Chen, L. Bao, Z. Zhang, and J. Yuan, "Joint hand-object 3d reconstruction from a single image with cross-branch feature fusion," *arXiv preprint* arXiv:2006.15561, 2020.

[79] M. Rami Koujan, A. Roussos, and S. Zafeiriou, "Deepfaceflow: In-the-wild dense 3d facial motion estimation," *arXiv preprint* 2005.07298, 2020.

[80] D. Thomas, "Real-time simultaneous 3d head modeling and facial motion capture with an rgb-d camera," *arXiv preprint* arXiv:2004.10557, 2020.

[81] C. Xin, Y. Li, X. Luo, T. Shao, J. Yu, K. Zhou, and Y. Zheng, "Autosweep: Recovering 3d editable objects from a single photograph," *IEEE Transactions on Visualization and Computer Graphics*, 2018.

[82] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, "Sketch2phc: Internet image montage," *ACM Transactions on Graphics (TOG)*, vol. 28, no. 5, pp. 1–10, 2009.

[83] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.

[84] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzarc, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8798–8807.

[85] T. Portenier, Q. Hu, A. Szabo, S. A. Bigdeli, P. Favaro, and M. Zwicker, "Faceshop: Deep sketch-based face image editing," *arXiv preprint* arXiv:1804.08972, 2018.

[86] S.-Y. Chen, W. Su, L. Gao, S. Xia, and H. Fu, "Deep generation of face images from sketches," *arXiv preprint* arXiv:2006.01047, 2020.

[87] X. Han, C. Gao, and Y. Yu, "Deepsketch2face: A deep learning based sketching system for 3d face and caricature modeling," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–12, 2017.

[88] A. Akman, Y. Sahillioglu, and T. M. Sezgin, "Generation of 3d human models and animations using simple sketches," 2020.

[89] X. Han, K. Hou, D. Du, Y. Qiu, S. Cui, K. Zhou, and Y. Yu, "Caricatureshop: Personalized and photorealistic caricature sketching," *IEEE Transactions on Visualization and Computer Graphics*, 2018.

[90] Y. Shen, C. Zhang, H. Fu, K. Zhou, and Y. Zheng, "Deepsketchhair: Deep sketch-based 3d hair modeling," *arXiv preprint* arXiv:1908.07198, 2019.

Yiwei Jin
Department of Computer Science
Zhejiang University
Hangzhou
China
E-mail address: jyw0506@zju.edu.cn

Diqiong Jiang
Department of Computer Science
Zhejiang University
Hangzhou
China
E-mail address: jdq1994@zju.edu.cn

Ming Cai
Department of Computer Science
Zhejiang University
Hangzhou
China
E-mail address: cm@zju.edu.cn

Received September 10, 2020