2024 IEEE International Conference on Robotics and Automation (ICRA)
May 13-17, 2024. Yokohama, Japan

Morphable-SfS: Enhancing Shape-from-Silhouette Via Morphable Modeling

Guoyu Lu

2024 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3503-8457-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICRA57147.2024.10610152

***Abstract*— Reconstructing accurate object shapes based on single image inputs is still a critical and challenging task, mainly due to the potential shape ambiguity and occlusion. Most existing single image 3D reconstruction approaches, either trained on stereo setting or structure-from-motion, estimate 2.5D visible models which generally reconstruct one viewpoint of objects. We propose a method to leverage both the general Morphable Model on common objects and a multi-view synthesis-based shape-from-silhouette model to reconstruct complete object shapes. We use the proposed method to exploit strong geometric and perceptual cues in 3D shape reconstruction. During the inference, the trained model is able to produce high-quality and complete meshes with finely detailed structures from a 2D image captured from arbitrary perspectives. The proposed method is evaluated on both large-scale synthetic ShapeNet and real-world Pascal 3D+ and Pix3D datasets. The proposed work achieves state-of-the-art results compared with other recent self-supervised methods. Moreover, it shows a good capability of being applied in the unseen object reconstruction tasks.**

I. Introduction

Accurate and swift 3D object reconstruction is pivotal for numerous tasks in computer vision and robotics, such as 3D model matching [8], manipulation [16], and understanding [46]. While multi-view reconstructions exploit geometric cues as prior information for unambiguous 3D model inference, deriving shape data solely from a single-view image remains a formidable challenge.

The recent development in deep learning has enabled researchers to focus more on applying deep networks to perform single-view 3D reconstruction [38][40][43]. Most works rely on volumetric structures to represent the shape and then conduct 3D convolution for reconstruction. However, such methods are largely limited due to the reconstruction quality of the low-resolution voxel grid, and the high expense on computational complexity and memory. Conventional techniques, such as SfM/SLAM, that depend on stereo and sequential images often necessitate intricate camera configurations and extensive camera motions. These methods can still face challenges from occlusions and produce incomplete 3D reconstructions. Following the approaches in [38], [40], we utilize polygon meshes with triangular faces for precise reconstruction. Distinctly, we integrate the 3D morphable model, commonly used in face modeling, into our deep network for general object reconstruction. This is iteratively refined using a novel-view synthesis-based deep shape-from-silhouette network, facilitating precise and efficient 3D reconstructions in a self-supervised fashion.

We present a cutting-edge deep learning architecture, **Morphable-SfS**, designed to reconstruct a 3D mesh from a singular image by perpetually updating the coefficients of the blend-shape basis. Upon shape prediction, it is revolved across varied perspectives in the 3D domain, synchronized with the given camera viewpoints. These 3D shapes, when associated with different perspectives, are cast into 2D color images and silhouette masks via a differentiable rendering network, fortified with a projection layer. This process is navigated by both multi-view silhouette and pixel color consistency. The ensuing reduction in shape ambiguity is largely attributed to the synergy between the morphable model prior and the multi-view rendering constraints. Given that the rendering loss is predominantly assessed in the projected image domain and falls short in depth and geometric detail constraints, we augment our method with 3D keypoint consistency across diverse viewpoints. The inherent nature of the morphable model facilitates the holistic recovery of the 3D shape, but with limited detailed constraints. Contrarily, shape-from-silhouette possesses the prowess to delve into intricate shapes and textures. Our intent is to harmonize the strengths of both by integrating SfS into the morphable model estimation network's framework. A visual depiction of our proposed architecture is provided in Fig. 1.

The salient contributions of our work are: 1. Harnessing the prowess of the morphable model, predominantly used in human face modeling, and embedding it in general object reconstruction with our shape estimation algorithm. This showcases the efficacy of a learning-based morphable model as a robust shape prior for general object reconstruction. 2. Unveiling an image rendering network tasked with predicting both object mask and color image from the modified 3D shapes. The enforcement of multi-view rendering consistencies, coupled with incessant updates from the morphable model network, enables the procurement of a comprehensive 3D shape without the crutch of ground-truth 3D data. 3. Beyond the confines of rendering loss, we incorporate multi-view 3D keypoint consistency to curtail shape ambiguity during the training phase.

II. Related work

A plethora of studies have delved into 3D object reconstruction and modeling, leveraging either a single image or multi-view images. In this context, we sequentially elucidate the 3D morphable model, the learning-centric 3D shape generation, and the neural rendering process.

3D Morphable Model has been extensively used for modeling 3D human faces [5]. While some approaches focus on single images [9], [36], others harness multiple frames [3] or unstructured photo sets [31]. These models utilize 3D parametric identities and principal components for geometry and associated expressions. Variations across

Guoyu Lu is with Intelligent Vision and Sensing Lab, University of Georgia, USA guoyu.lu@uga.edu

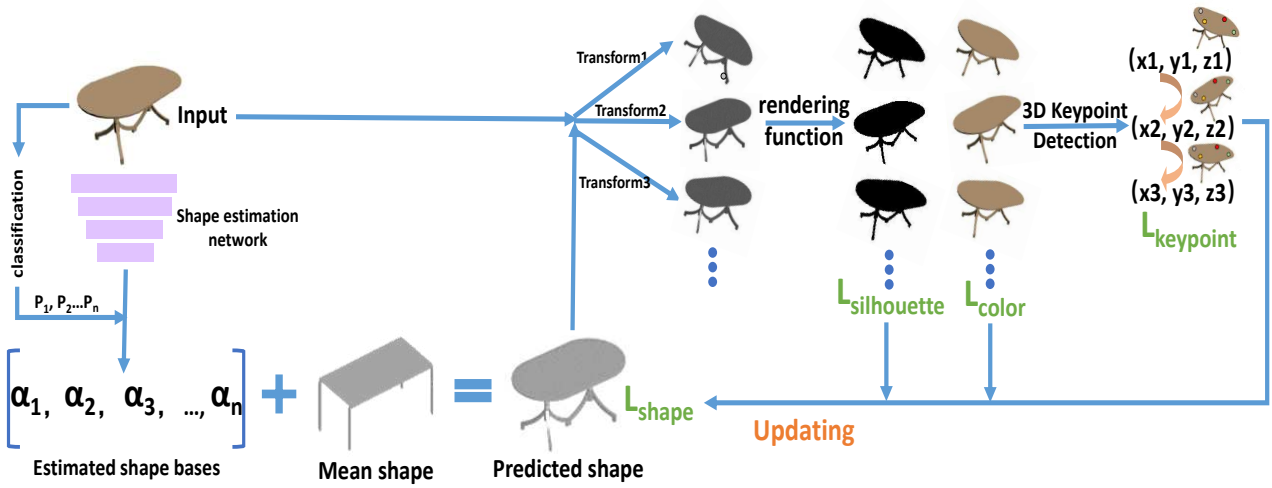

Fig. 1. An illustration of our proposed deep Morphable-SfS network for 3D shape reconstruction. Object-level shape reconstruction is achieved by continuously updating the blend-shape basis coefficients, based on multi-view rendering (silhouette and color) and keypoint consistencies in the Shape-from-silhouette network.

different individuals are captured using principal components derived from laser scans in a reduced-dimensional subspace [4]. Traditional optimization strategies leverage personalized models [6], [12], [18] to deduce 3D shapes considering texture or lighting. Contemporary research has seen a shift towards learning-based methodologies that directly infer 3D faces from an image [21], [30], [32], [11]. While the majority of such research targets human face modeling, comprehensive 3D shape reconstructions like the human head are less explored [27]. Leveraging the inherent advantage of the 3D morphable model to recover a holistic 3D shape, we propose its integration into our general object reconstruction endeavor. Given the inherent capability of the 3D morphable model to holistically recover an initial 3D shape and align with 2D image observations, we advocate its integration into our broader object reconstruction framework.

Single Image Shape Generation, a burgeoning area of research, strives to reconstruct 3D shapes from singular RGB images [38], [15], [39], [45], depth images [44], [48], or a combination thereof [10]. Approaches span from alignment-based [20], [2], deformation-based [40], [23], to direct regression using convolutional neural networks [42], [13]. The 3D Recurrent Reconstruction Neural Network, DIRT [17], was inspired by traditional LSTM [34] to convert single or multi-view object images to 3D shapes. However, these techniques often overlook the potential of morphable geometric cues for shape constraint enhancement. Also, integrating class prediction, rendering networks, and multi-view consistencies alleviates the need for intensive 3D supervision.

Neural Rendering Methods: Rendering traditionally involves projecting a 3D shape onto a 2D plane. While early methods used derivatives of rendering to link 2D image variations with 3D modifications [14], [47], others explored renderings of voxel grids [29], [37], point clouds [19], [1], meshes [22], [24], and implicit surfaces [28]. Often, these approaches utilized differentiable functions and approximations, trading off image clarity and 3D shape fidelity. Recent endeavors [22], [35] have incorporated these into 3D reconstruction networks, placing supervisions solely on the 2D images. While many methods leverage encoder-decoder architectures for projection, our approach distinguishes itself by deriving shape geometry from a novel deep morphable model network and sourcing texture colors from a dedicated color-rendering CNN. Rendering is executed using a Soft Rasterizer [25], and optimized via multi-view perceptual, silhouette, and keypoint consistencies.

III. Morphable-SfS for Shape Reconstruction

3D shape estimation from singular images is non-trivial due to unknown poses and occlusions. We aim for continual geometry updates through deep morphable models and multi-view constraints.

A. Deformable Deep 3D Morphable Model

The 3D Morphable Model (3DMM) in our framework serves a dual purpose: it offers robust shape priors and facilitates iterative optimization. Such a design assists in circumventing challenges such as incompleteness, occlusion, and ambiguity inherent when learning shapes from a solitary view during inference. The geometry of a typical object having n vertices and encompassing the shape coordinates (x, y, z) can be articulated via a shape matrix:

$$S = \begin{bmatrix} x_1, & x_2, & \dots & , x_n \\ y_1, & y_2, & \dots & , y_n \\ z_1, & z_2, & \dots & , z_n \end{bmatrix}_{3\times n} \quad (1)$$

Here, the object shape S comprises x, y, and z coordinates, sized $3\times n$ (with $n = 642$). To streamline the training for our shape estimation network, we commence with object classification, deriving probability scores for sub-classes (e.g., suv, sedan, coupe, van) within primary categories (e.g., car, chair, table). For each sub-class, we systematically reduce edges in the 3D model, aiming to minimize surface quadric errors until attaining a uniform count of vertices and mesh faces suitable for principal component analysis (PCA). The 3D shape S undergoes updates by morphing an average model using a sub-class weighted amalgamation of a set of trainable parameters. These parameters resonate with the coefficients of the principal shape basis, leading to:

$$M_c + \sum_i P_i \cdot U_{i,s} \cdot diag(\sigma_{i,s}) \cdot \alpha_{i,s} \rightarrow S \tag{2}$$

Given $M_c \in \mathbb{R}^{3n}$ as the mean object shape for a specific category c, and denoting $U_{i,s}$ as the matrix of principal components for each shape basis i — where each column vector is given by $u_{i,j}$ and $k \ll n$ — we can define $\sigma_{i,s}$ as the standard deviation. The shape can then be parameterized and subsequently updated using shape vectors $\alpha_{i,s}$.

For any given input image, we determine the corresponding low-dimensional shape basis of the selected 3D shapes, each signifying a different sub-class from ShapeNet. We retain only the initial 20 dimensions of the shape vectors, utilizing them as a shape basis for every individual sub-class. The final resultant shape emerges from an incessantly updated fusion of the mean model shape and the blended shape basis. This fusion is influenced by the trainable coefficient parameters and the classification scores. An exemplary representation of our proposed trainable and modifiable 3D morphable model is depicted in Fig. 2. To counteract extreme vertex movement in the reconstructions, we integrate a Laplacian loss L_{shape}, mirroring the approach adopted in [38] and [25].

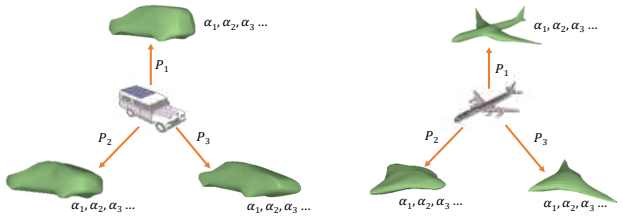

Fig. 2. An illustration of our 3D morphable model. For each sub-class, we continuously update a group of trainable principal coefficients for reconstruction.

B. Differentiable Neural Rendering

Given a reconstructed 3D mesh, let's denote its vertices as $v_1, v_2, \ldots, v_n$ and corresponding faces as $f_1, f_2, \ldots, f_N$. Here, v_i defines the 3D position of the i^{th} vertex, while f_j represents the j^{th} index of the triangle face, composed of three vertices. The process of rendering a 3D object entails transforming these vertices from the object space to an image plane. Rather than discretely sampling feature points, which results in a non-differentiable operation or making use of approximate derivatives as in [41], we leverage the aggregate function detailed in [25]. This allows us to compute the probability map P_j for every pixel i and a relative depth for 3D points, d_j. Consequently, both the rendered color image I_{color} and the silhouette mask $I_{silhouette}$ for each pixel i can be articulated by integrating the object color C_j with the background color C_b as:

$$\begin{aligned} I^i_{color} &= \sum_j w^i_j C^i_j + w^i_b C_b \\ I^i_{silhouette} &= 1 - \prod_j (1 - P^i_j) \end{aligned} \tag{3}$$

where w_j and w_b weight the foreground object color C_j and background color C_b, respectively, and are defined as $w^i_j = \frac{P^i_j exp(d^i_j/\gamma)}{\sum_k P^i_k exp(d^i_k/\gamma) + exp(\epsilon/\gamma)}$ and $\sum_j w^i_j + w^i_b = 1$. ϵ and γ are both small constants of 1e-4.

To estimate the color maps C, we employ a color generation network designed to predict color values. Given an RGB image, a pre-trained ResNet-18 network serves as a feature extractor, producing 512-dimensional global features. The network's output consists of $n \times 3$ color values, derived from the FC layers, where n represents the number of sampled color classes. These values are then incorporated into the aggregation function and are pivotal in minimizing the rendering loss between the generated and actual images.

C. Multi-view Rendering and Geometric Constraints

Leveraging the aforementioned deep morphable model for 3D shape generation and the differentiable rendering process for image mapping, we suggest rotating the produced mesh to various perspectives, as illustrated in Fig. 3. In practice, numerous views of each scene are captured, and a subset is randomly sampled during each training iteration. This approach facilitates the development of models robust to a wide range of camera transformations. The multi-view silhouettes impose stringent constraints on the global geometry of our reconstruction model. The transformation of the predicted 3D shape derived from an input image I_{input} through the deep 3D morphable model Morph$(\cdot)$ to a rendered image from a distinct viewpoint I_{render}, using the transformation $\mathbf{T}$ and the aforementioned neural rendering function $R(\cdot)$, can be described as:

$$I_{render} = R(\boldsymbol{T} \cdot Morph(I_{input})) \tag{4}$$

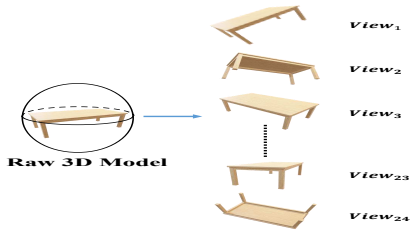

Fig. 3. Illustration of the multi-view generation. Given a generated 3D object, we generate 24 random views (12 horizontal and 12 vertical directions).

Utilizing the aforementioned mapping, our comprehensive network perpetually refines the parameters of the adaptable shape coefficients, mapping a singular 2D image to a 3D mesh. We aim for the rendered color image and the object silhouette mask derived from the produced object to achieve the highest degree of photo-realism. The multi-view rendering losses for both silhouette and color images are crafted to diminish discrepancies between the synthesized renderings and the authentic color and silhouette representations. We incorporate a blend of SSIM and perceptual loss, denoted by $\Phi_{\text{vgg}}(\cdot)$, as given:

$$\begin{aligned} L_{rendering} = \lambda \cdot [(1 - SSIM(I_{render}, I_{real})) + \\ (1 - SSIM(M_{render}, M_{real}))] + \\ (1 - \lambda) \cdot [||\Phi_{vgg}(I_{render}) - \Phi_{vgg}(I_{real})||_1 + \\ ||\Phi_{vgg}(M_{render}) - \Phi_{vgg}(M_{real})||_1] \end{aligned} \tag{5}$$

Given that I_{render} and M_{render} denote the rendered color image and silhouette mask in a batch, respectively, the function

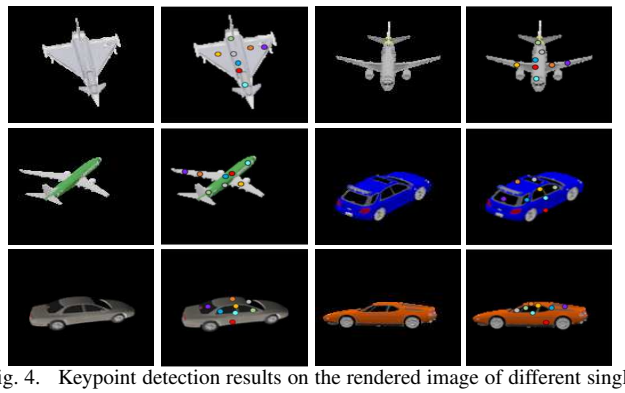

Fig. 4. Keypoint detection results on the rendered image of different single objects from the test split of the ShapeNet dataset.

$\Phi_{\text{vgg}}(\cdot)$ fetches feature vectors from the layers 'conv1-2', 'conv2-2', 'conv3-2', 'conv4-2', and 'conv5-2' of the VGG-19 network, which is pretrained on the ImageNet dataset. In contrast to utilizing $L1$ or $L2$ for rendering, the integration of SSIM and perceptual loss results in more defined images.

While the rendering loss does not impose geometric constraints on the reconstructed shape, we introduce spatial correspondence using consistent 3D keypoints across multiple views of the same object instance. A dedicated keypoint detection network predicts 3D keypoints from an input 2D image without reliance on ground truth keypoints for supervision. With known relative camera poses, this network processes each view through 13 stacked dilated convolution layers (with 64 channels), culminating in the output of eight 3D keypoints in the form of pixel coordinates xy and depth z. In inference, it gleans 3D keypoints directly from a single image, eschewing the need for pose data. The overarching objective of this geometric constraint is to align the projected 3D keypoints on the initial image with their corresponding positions in subsequent views. Should the predicted keypoint in 3D space from the rendered image I_{render} be $[x^r, y^r, z^r]$, and using the projection function $\pi(\cdot)$, the corresponding 2D keypoint in image coordinates is denoted as $[u^r, v^r]$. The re-projected 2D keypoints from the other $M-1$ viewpoints are thus expressed as $[\tilde{u^r}, \tilde{v^r}] = \pi \times \mathbf{P} \pi^{-1}[u_j^r, v_j^r]$. Consequently, the multi-view keypoint consistency loss L_{consis} is:

$$L_{consis} = \frac{1}{N(M-1)} \sum_{i=1}^{N} \sum_{j=1}^{M-1} ||(u_i^r, v_i^r) - (\tilde{u_j^r}, \tilde{v_j^r})||^2 \quad (6)$$

To guarantee that all estimated 2D keypoints fall within the object's silhouette regions, we apply an additional regional constraining loss L_{region} to enforce the relationship between the pixel intensity of keypoints and the silhouette mask:

$$L_{region} = \frac{1}{N} \sum_{i=1}^{N} [-log \sum_{u,v} M(u,v) I_i(u,v)] \quad (7)$$

Ensuring that all predicted 2D keypoints reside within the object silhouette regions is vital. Hence, we introduce a regional constraining loss L_{region} which enforces the pixel intensity of the keypoint and the silhouette mask to adhere to the relationship:

$$L_{sep} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{i \neq j}^{N} max(0, c - ||(u_i^r, v_i^r) - (u_j^r, v_j^r)||^2) \quad (8)$$

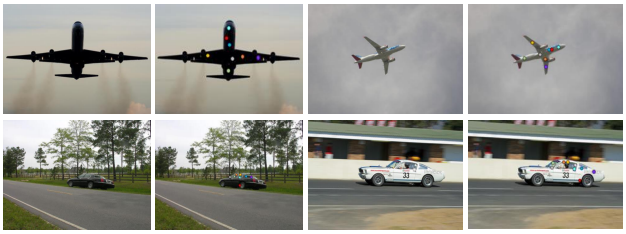

Fig. 5. Generalization of keypoint detection on real images from the real-world PASCAL 3D+ dataset.

Category	DIRT [17]	Pix2Mesh [38]	SoftRas [25]	Learn3D [26]	Ours
airplane	42.6 (-)	31.4 (2.82)	50.9 (1.96)	**56.5** (**1.62**)	53.4 (1.78)
car	66.1 (-)	55.2 (2.88)	67.2 (**2.03**)	67.0 (2.31)	**67.6** (2.16)
chair	43.9 (-)	50.7 (2.76)	41.9 (7.09)	41.5 (8.01)	**43.4** (**5.97**)
table	42.0 (-)	40.9 (5.47)	38.0 (6.90)	41.7 (6.92)	**45.1** (**5.59**)
display	44.0 (-)	45.8 (3.02)	47.8 (3.73)	**52.8** (**3.60**)	51.9 (3.91)
sofa	62.6 (-)	61.3 (2.39)	55.9 (2.65)	59.2 (2.27)	**63.1** (**2.23**)
lamp	28.1 (-)	32.3 (9.79)	32.7 (9.84)	37.8 (8.67)	**38.0** (**8.43**)

TABLE I

IOU (AND CD) RESULTS ON THE SHAPENET DATASET, COMPARED WITH RECENT METHODS WITH [17][38] AND WITHOUT [25][26] DIRECT 3D SUPERVISION. FOR IOU, THE HIGHER VALUE MEANS THE BETTER RECONSTRUCTION. FOR CD, THE LOWER VALUES INDICATE THE BETTER PERFORMANCE. BOLD INDICATES THE BEST RESULTS AMONG THE METHODS WITHOUT DIRECT SUPERVISION, AND THE UNDERSCORE INDICATES THE BEST RESULTS AMONG ALL THE COMPARED METHODS.

Therefore, the comprehensive spatial consistency loss L_{spatial} for geometric constraints is an amalgamation of L_{consis}, L_{region}, and L_{sep}.

Loss for Joint Training: In pursuit of a refined object shape, we amalgamate multi-view rendering with geometric constraints encompassing shape geometry and texture into an end-to-end schema. By leveraging this joint loss function, we harness insights from both avenues, enhancing performance and tackling shape ambiguities and truncations. Formally, the culminating objective is framed as a weighted aggregation of the three constituent losses, assigned weights of 0.005, 1.0, and 0.1 respectively.

$$L_{Ge3DMM} = \lambda_{shape} L_{shape} + \lambda_{rendering} L_{rendering} + \lambda_{spatial} L_{spatial} \quad (9)$$

IV. EXPERIMENTS

Training Data: Our dataset is sourced from ShapeNet [7], a comprehensive repository boasting around 51K 3D models spread over 13 primary categories. We adopt a distinct training/testing partition for diverse object classes in alignment with the protocol in [17], which encompasses cars, chairs, planes, monitors, and more. Each CAD model is normalized such that its largest dimension spans the interval $(-1, 1)$. We render these models from 24 eclectic vantage points, split evenly between horizontal and vertical orientations.

In addition to the ShapeNet collection, the PASCAL 3D+ dataset augments our data. We chiefly cull the car, airplane, and sofa categories from PASCAL 3D+, mirroring those in ShapeNet, to scrutinize our technique across varied scenarios. Given that ShapeNet is replete with rendered imagery, we employ the Pix3D [33] dataset for our test segment, facilitating evaluation of our approach against genuine images coupled with their 3D counterparts.

Experimental Settings: Our training strategy entails a three-pronged approach, harnessing the capabilities of a 3D

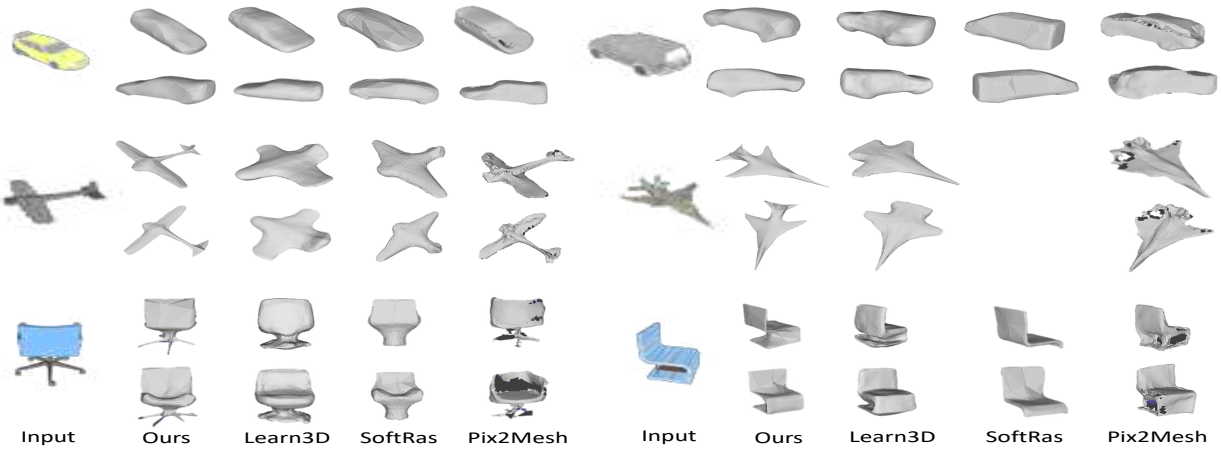

Fig. 6. Comparison of reconstruction performance on ShapeNet dataset (shown in two views). Left to right: raw input image; our reconstructed shape; reconstruction from [26]; reconstruction from [25]; reconstruction from [38]. Empty space represents that the corresponding method fails to reconstruct a 3D shape based on the input image.

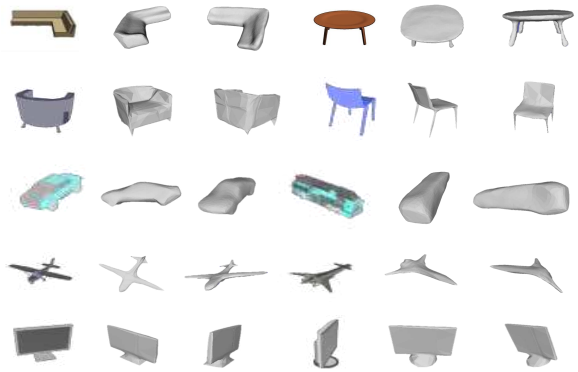

Fig. 7. Sample images and reconstructed shapes from ShapeNet and Pascal 3D+ datasets. Left to right for each sample: Input 2D image; 3D shape from two different viewpoints.

Fig. 8. Generalization results on the real-world Pix3D dataset.

morphable network, a differentiable rendering network, and a 3D keypoint detection network. Importantly, our approach is self-supervised, devoid of any 3D oversight.

Both the generated mean shape and the reconstructed shape share a consistent mesh structure, comprising 642 vertices and 1280 faces. During testing, our method requires merely a single input image to furnish the predicted 3D object and the associated 3D keypoints.

We rigorously benchmark our methodology against contemporaneous approaches reliant on either direct 3D supervision or 2D images for shape synthesis. This includes the works of [38], [25], and [26]. For a comprehensive insight, we direct the reader to the extensive visual and quantitative comparisons presented in Table I and Figs. 6-10.

Fig. 9. Our reconstruction pipeline is able to recover the texture of the 3D model. For each sample, the left side is the input image, and the right side is the projected image from our reconstructed colored meshes.

A. Keypoint Detection Results

Our self-supervised general-object keypoint detection is first evaluated qualitatively on classes such as cars, chairs, monitors, and planes from the ShapeNet dataset's test split. Fig. 4 showcases randomly selected objects across various viewpoints. Despite the absence of annotated keypoints for direct supervision, our network demonstrates a commendable ability to infer stable and precise keypoint locations from a single rendered input image.

The network's generalization capability is further exemplified in Fig. 5, where we examine its performance in real-world scenarios using the PASCAL 3D+ dataset, specifically on airplane and car scenes. Notably, even without training on this dataset, our model produces consistent and meaningful keypoint inferences.

B. Shape Reconstruction from A Single Image

For a comprehensive assessment of the shape and point distribution, we perform quantitative comparisons on 3D shape reconstruction using both IoU and chamfer distance (CD) metrics, as presented in Table I. Our model outperforms other state-of-the-art methods without direct 3D supervision in most categories, such as car, chair, table, sofa, and lamp. Remarkably, our approach even surpasses some supervised methods in categories like car, table, sofa, and lamp. This performance underscores the efficacy of our introduced 3D morphable network with multi-view rendering and geometric constraints. Specifically, our method consistently demonstrates an increase in IoU and a decrease in CD metrics against the evaluated ground truth shape. After training, our model can accurately and comprehensively reconstruct a 3D shape from a single 2D image across various categories and

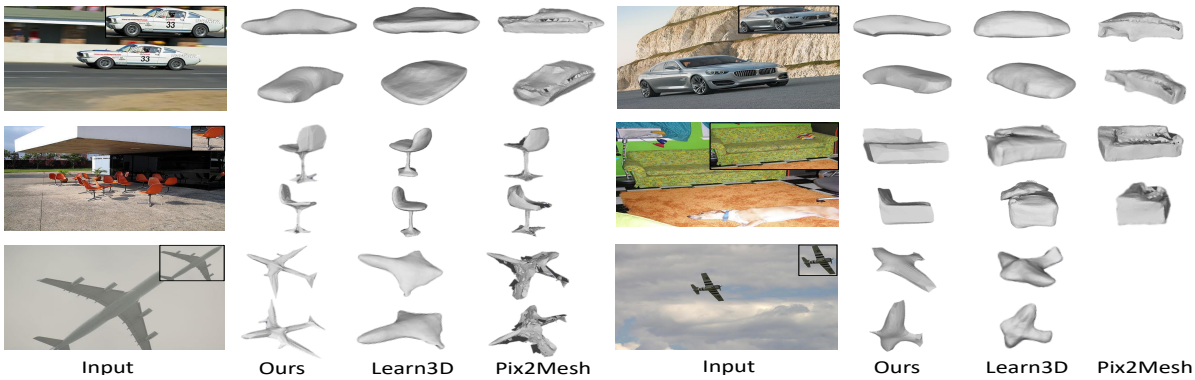

Fig. 10. Comparison of reconstruction performance on PASCAL 3D+ dataset (shown in two views). Left to right: raw input image (detected regions as input); our reconstructed shape; reconstruction from [26]; reconstruction from [38]. Empty space represents that the corresponding method fails to reconstruct a 3D shape based on the input image. Reconstructions are directly from the pre-trained model on ShapeNet without re-training or fine-tuning.

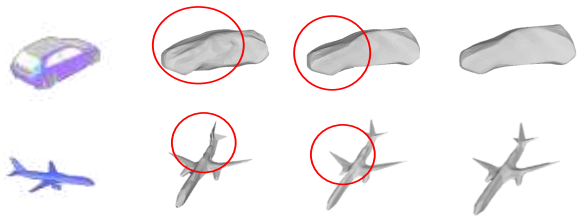

Fig. 11. Ablation study of reconstruction results on ShapeNet with different baseline methods and our full pipeline. Left to right: input image; reconstruction without morphable model; reconstruction without keypoint constraint; our full reconstruction.

viewpoints, as illustrated in Fig. 7.

Visual predictions and comparisons on the ShapeNet and PASCAL 3D+ datasets are depicted in Fig. 6 and Fig. 10. From [38], one can notice distortions and numerous holes when viewed from different perspectives, as well as missing structures, like the bottom of chairs or car wheels, as seen in [25]. Although [26] excels in comparison to others, it doesn't adequately constrain details, leading to smoothed approximations. Our method, with its multi-view constraints, closely aligns with the ground truth shape. Fig. 8 and Fig. 10 display the robustness of our model in real-world settings (note that for complex scenes, we detect and resize objects as inputs across all methods). While meshes are shown without textures for fair comparisons, our approach can reconstruct not just the shape but also color closely resembling the input 2D image. However, training was restricted to just 50 distinct colors. The input image alongside the projected image from the inferred 3D model is exhibited in Fig. 9.

C. Ablation Study and Analysis

The ablation study, comprising both quantitative and qualitative results from the ShapeNet dataset, is presented in Table II and Fig. 11. As seen in Table II, our proposed approach outperforms various baselines—whether replacing the morphable model with a generic 3D encoder-decoder, omitting multi-view keypoint cues, or solely using SSIM loss in $L_{rendering}$. This underscores the efficacy of our comprehensive training methodology. Specifically, our method elevates the IoU score by 4.7 and 2.9 points over baselines lacking the morphable model and keypoint modules, respectively. Fig. 11 further highlights the advantage of our full pipeline: while baselines may exhibit irregular surfaces (as in cars) or partial omissions (as in airplanes), our method

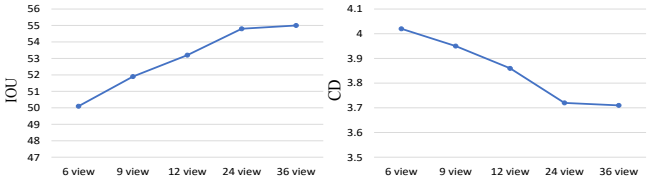

Fig. 12. Ablation study of how the number of viewpoints affect the reconstructed shape geometry (in IoU and CD) and texture (in PSNR).

mitigates these issues.

Moreover, our Morphable-SfS demonstrates an iterative refinement in the reconstruction of both shape geometry and color texture, contingent on the number of rendered viewpoints. As illustrated in Fig. 12, the IoU performance initially ascends as CD significantly drops, but plateaus beyond a certain number of views, indicating a subsequent rise in computational demands.

	airplane	chair	car	mean
Ours **w/o** morphable model	49.7	40.2	62.9	50.1
Ours **w/o** keypoints	51.2	40.9	65.3	51.9
Ours **w/o** perceptual loss in $L_{rendering}$	52.7	41.9	66.3	53.9
Our full pipeline	**53.4**	**43.4**	**67.6**	**54.8**

TABLE II

ABLATION STUDY OF OUR MODEL TRAINED WITHOUT SPECIFIC COMPONENT ON IOU RESULTS.

V. Conclusion

In this study, we introduce *Morphable-SfS*: a versatile, deep, morphable model-driven shape reconstruction network fortified with multi-view rendering and geometric constraints. Our methodology enhances the reconstruction process by ensuring that rendering outputs align with the initial inputs. Additionally, we incorporate a self-supervised 3D keypoint detection network to refine the reconstruction against sparse keypoint geometries. Coupled with *Morphable-SfS*, our learning-based neural rendering facilitates the derivation of high-caliber textured images, superseding traditional non-differentiable variants. Through extensive 3D modeling experiments, our approach consistently outperforms recent state-of-the-art techniques across both rendered and real-world datasets. Furthermore, ablation analyses of various components and configurations within our workflow substantiate the efficacy of our proposed paradigm.

Acknowledgement: This publication is based upon work supported by NSF under Awards No. 2334690 and 2334624.

REFERENCES

[1] Kara-Ali Aliev, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. *arXiv preprint arXiv:1906.08240*, 2(3):4, 2019.

[2] Mohamed El Banani, Jason J Corso, and David F Fouhey. Novel object viewpoint estimation through reconstruction alignment. In *CVPR*, pages 3113–3122, 2020.

[3] Christian Baumberger, Mauricio Reyes, Mihai Constantinescu, Radu Olariu, Edilson de Aguiar, and Thiago Oliveira Santos. 3d face reconstruction from video using 3d morphable model and silhouette. In *SIBGRAPI 2014*, pages 1–8. IEEE, 2014.

[4] Volker Blanz, Curzio Basso, Tomaso Poggio, and Thomas Vetter. Reanimating faces in images and video. In *CGF*, volume 22, pages 641–650. Wiley Online Library, 2003.

[5] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[6] Chen Cao, Hongzhi Wu, Yanlin Weng, Tianjia Shao, and Kun Zhou. Real-time facial animation with image-based dynamic avatars. *TOG*, 35(4), 2016.

[7] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[8] Haowen Deng, Tolga Birdal, and Slobodan Ilic. Ppfnet: Global context aware local features for robust 3d point matching. In *CVPR*, pages 195–205, 2018.

[9] Claudio Ferrari, Giuseppe Lisanti, Stefano Berretti, and Alberto Del Bimbo. Dictionary learning based 3d morphable model construction for face recognition with varying expression and pose. In *3DV 2015*, pages 509–517. IEEE, 2015.

[10] Sam Fowler, Hansung Kim, and Adrian Hilton. Towards complete scene reconstruction from single-view depth and human motion. In *BMVC*, 2017.

[11] Zhongpai Gao, Juyong Zhang, Yudong Guo, Chao Ma, Guangtao Zhai, and Xiaokang Yang. Semi-supervised 3d face representation learning from unconstrained photo collections. In *CVPR*, pages 348–349, 2020.

[12] Pablo Garrido, Levi Valgaerts, Chenglei Wu, and Christian Theobalt. Reconstructing detailed dynamic face geometry from monocular video. *TOG*, 32(6):158–1, 2013.

[13] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *CVPR*, pages 10833–10842, 2019.

[14] Ioannis Gkioulekas, Anat Levin, and Todd Zickler. An evaluation of computational imaging techniques for heterogeneous inverse scattering. In *ECCV*, pages 685–701. Springer, 2016.

[15] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *CVPR*, pages 216–224, 2018.

[16] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, pages 11807–11816, 2019.

[17] Paul Henderson and Vittorio Ferrari. Learning single-image 3D reconstruction by generative modelling of shape, pose and shading. *IJCV*, 2019.

[18] Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. Dynamic 3d avatar creation from hand-held video input. *ToG*, 34(4):1–14, 2015.

[19] Eldar Insafutdinov and Alexey Dosovitskiy. Unsupervised learning of shape and pose with differentiable point clouds. *arXiv preprint arXiv:1810.09381*, 2018.

[20] Hamid Izadinia, Qi Shan, and Steven M Seitz. Im2cad. In *CVPR*, pages 5134–5143, 2017.

[21] Amin Jourabloo and Xiaoming Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *CVPR*, pages 4188–4196, 2016.

[22] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *CVPR*, pages 3907–3916, 2018.

[23] Johannes Kepler. Learning free-form deformations for 3d object reconstruction. *Deep Learning Approaches for 3D Inference from Monocular Vision*, page 41, 2020.

[24] Shichen Liu, Weikai Chen, Tianye Li, and Hao Li. Soft rasterizer: Differentiable rendering for unsupervised single-view mesh reconstruction. *arXiv preprint arXiv:1901.05567*, 2019.

[25] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. A general differentiable mesh renderer for image-based 3d reasoning. *TPAMI*, 44(1):50–62, 2020.

[26] Bo Peng, Wei Wang, Jing Dong, and Tieniu Tan. Learning pose-invariant 3d object reconstruction from single-view images. *Neurocomputing*, 423:407–418, 2021.

[27] Stylianos Ploumpis, Evangelos Ververas, Eimear O'Sullivan, Stylianos Moschoglou, Haoyang Wang, Nick Pears, William Smith, Baris Gecer, and Stefanos P Zafeiriou. Towards a complete 3d morphable model of the human head. *TPAMI*, 2020.

[28] Edoardo Remelli, Artem Lukoianov, Stephan R Richter, Benoît Guillard, Timur Bagautdinov, Pierre Baque, and Pascal Fua. Meshsdf: Differentiable iso-surface extraction. *arXiv preprint arXiv:2006.03997*, 2020.

[29] Danilo Jimenez Rezende, SM Eslami, Shakir Mohamed, Peter Battaglia, Max Jaderberg, and Nicolas Heess. Unsupervised learning of 3d structure from images. *arXiv preprint arXiv:1607.00662*, 2016.

[30] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *CVPR*, pages 1259–1268, 2017.

[31] Joseph Roth, Yiying Tong, and Xiaoming Liu. Adaptive 3d face reconstruction from unconstrained photo collections. In *CVPR*, pages 4197–4206, 2016.

[32] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *CVPR*, pages 7763–7772, 2019.

[33] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *CVPR*, pages 2974–2983, 2018.

[34] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*, 2014.

[35] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. State of the art on neural rendering. In *CGF*, volume 39, pages 701–727. Wiley Online Library, 2020.

[36] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *CVPR*, pages 7346–7355, 2018.

[37] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *CVPR*, pages 2626–2634, 2017.

[38] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Hang Yu, Wei Liu, Xiangyang Xue, and Yu-Gang Jiang. Pixel2mesh: 3d mesh model generation via image guided deformation. *TPAMI*, 43(10):3600–3613, 2020.

[39] Peng-Shuai Wang, Chun-Yu Sun, Yang Liu, and Xin Tong. Adaptive o-cnn: a patch-based deep representation of 3d shapes. *TOG*, 37(6):1–11, 2018.

[40] Chao Wen, Yinda Zhang, Zhuwen Li, and Yanwei Fu. Pixel2mesh++: Multi-view 3d mesh generation via deformation. In *ICCVW*, pages 1042–1051, 2019.

[41] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *CVPR*, pages 7467–7477, 2020.

[42] Jiajun Wu, Tianfan Xue, Joseph J Lim, Yuandong Tian, Joshua B Tenenbaum, Antonio Torralba, and William T Freeman. Single image 3d interpreter network. In *ECCV*, pages 365–382. Springer, 2016.

[43] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *ICCVW*, pages 2690–2698, 2019.

[44] Bo Yang, Stefano Rosa, Andrew Markham, Niki Trigoni, and Hongkai Wen. Dense 3d object reconstruction from a single depth view. *TPAMI*, 41(12):2820–2834, 2018.

[45] Yuan Yao, Nico Schertler, Enrique Rosales, Helge Rhodin, Leonid Sigal, and Alla Sheffer. Front2back: Single view 3d shape reconstruction via front to back prediction. In *cvpr*, pages 531–540, 2020.

[46] Lintao Zheng, Chenyang Zhu, Jiazhao Zhang, Hang Zhao, Hui Huang, Matthias Niessner, and Kai Xu. Active scene understanding via online semantic reconstruction. In *CGF*, volume 38, pages 103–114. Wiley Online Library, 2019.

[47] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, pages 1851–1858, 2017.

[48] Chuhang Zou, Ersin Yumer, Jimei Yang, Duygu Ceylan, and Derek Hoiem. 3d-prnn: Generating shape primitives with recurrent neural networks. In *ICCV*, pages 900–909, 2017.