Self-Supervised 3D Mesh Reconstruction from Single Images

Tao Hu[1] Liwei Wang[1] Xiaogang Xu[1] Shu Liu[2] Jiaya Jia[1,2]
[1] The Chinese University of Hong Kong [2] SmartMore
{taohu,lwwang,xgxu,leojia}@cse.cuhk.edu.hk, sliu@smartmore.com

Abstract

Recent single-view 3D reconstruction methods reconstruct object's shape and texture from a single image with only 2D image-level annotation. However, without explicit 3D attribute-level supervision, it is still difficult to achieve satisfying reconstruction accuracy. In this paper, we propose a Self-supervised Mesh Reconstruction (SMR) approach to enhance 3D mesh attribute learning process. Our approach is motivated by observations that (1) 3D attributes from interpolation and prediction should be consistent, and (2) feature representation of landmarks from all images should be consistent. By only requiring silhouette mask annotation, our SMR can be trained in an end-to-end manner and generalizes to reconstruct natural objects of birds, cows, motorbikes, etc. Experiments demonstrate that our approach improves both 2D supervised and unsupervised 3D mesh reconstruction on multiple datasets. We also show that our model can be adapted to other image synthesis tasks, e.g., novel view generation, shape transfer, and texture transfer, with promising results. Our code is publicly available at https://github.com/Jia-Research-Lab.

1. Introduction

Single-view 3D Object Reconstruction is to recover 3D information, such as shape and texture, of the object from a single image [7, 15, 20, 41]. It is a long-standing problem in computer vision with various applications, including 3D scene analysis, robot navigation, and virtual/augmented reality. Traditional methods usually fit the parameters of a 3D prior morphable model, such as 3DMM [1] for faces and SMPL [23] for human. Building these prior models is expensive and time-consuming, and thus is not quickly applicable to many different natural objects.

In the deep learning era, deep models can learn to reconstruct 3D objects in a supervised manner [7]. 3D supervised reconstruction methods [4, 39, 5, 6, 25, 31, 11] directly minimize the discrepancy between the ground-truth 3D attributes and the predicted ones. They usually achieve supreme performances but have to be trained on synthesized or 3D scanned datasets with ground-truth 3D annotations. Meanwhile, since 2D attributes (*e.g.*, silhouette mask or landmark) are usually easier to be obtained than 3D attributes, 2D supervised reconstruction methods [16, 22, 3, 15, 28, 8] do not require 3D annotations. The key module of 2D supervised approaches is a differentiable render [24, 16, 22], which builds a differentiable stream to link 3D model space to 2D images and makes it possible to reconstruct 3D objects through 2D image-level supervision.

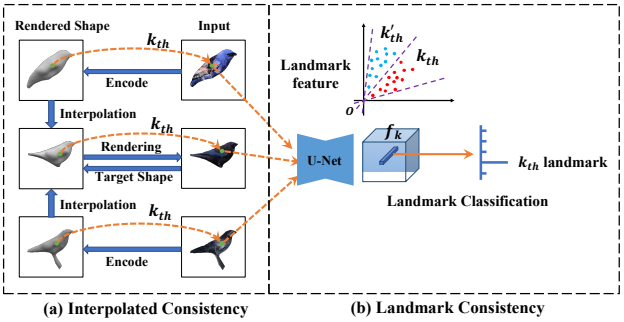

Figure 1: Our proposed self-supervised methods for 3D mesh reconstruction. Interpolated consistency provides fine-grained 3D annotations to train the reconstruction model by self-supervised regression. Landmark consistency further improves the reconstructed quality in local regions by self-supervised classification for landmarks.

Though 2D supervised reconstruction alleviates the dependency on 3D annotation, it is mainly to minimize the image-level reconstruction error and does not ensure 3D attribute prediction accuracy. The 3D reconstructed results provided in the work of ARCH [11] show that combining 2D with 3D supervision can further improve the accuracy of 3D reconstruction. Therefore, we raise the question *if it is possible to achieve 3D attribute-level reconstruction only with 2D annotation*.

In this work, we propose Self-Supervised Mesh Reconstruction (SMR) to reconstruct category-specific 3D mesh objects from single images. 3D attributes, including camera, shape, texture, and light, are first predicted by attribute encoder and then are supervised at both 2D image and 3D

attribute levels. At the 2D image level, similar to other 2D supervision approaches [3, 15], reconstructed models are rendered to the same images as original input. At the 3D attribute level, as illustrated in Fig. 1, our two novel self-supervised methods, *i.e*. Interpolated Consistency (IC) and Landmark Consistency (LC), further improve the learning process of 3D mesh attributes.

For Interpolated Consistency (IC), our motivation is that the interpolated 3D attributes should be consistent with their rendered images' encoded attributes. In other words, the interpolated attributes can be treated as the pseudo 3D annotation to train the reconstruction model by self-supervised learning, as illustrated in Fig. 1(a). Compared with the original [10, 18] or randomly augmented attribute in [30], our interpolated attributes can render images with more viewpoints, geometrical structures, and appearances, thus is more efficient to promote the learning process of the target attribute encoder.

Moreover, we propose Landmark Consistency (LC) to further improve landmark-level reconstruction, as illustrated in Fig. 1(b). If the local parts of a 3D object are well reconstructed, visible landmark feature should be consistent across all images. We treat the mesh vertices as the landmarks of objects. Then the feature of each visible landmark is classified to the mesh index. This ensures specialty of each landmark and improves the local quality of 3D mesh reconstruction.

Our final contributions are:

1. We propose interpolated consistency and landmark consistency as two self-supervised methods to learn the 3D mesh attributes.
2. We propose SMR to reconstruct category-specific 3D mesh objects from a collection of single images. It is an end-to-end training approach and is general to model 3D objects.
3. Experiments on the ShapeNet [2] and the BFM [41] datasets demonstrate that our method steadily improves both 2D supervised and unsupervised reconstruction. On the CUB-200-2011 [38] dataset, our SMR outperforms current state-of-the-art mesh reconstruction methods [15, 18, 20].

2. Related Work

According to the types of supervision, modern single-view 3D reconstruction can be mainly divided into three groups, *i.e*. 3D supervised [4, 39, 5, 6, 25, 31, 11], 2D supervised [16, 22, 3, 15, 28, 8], and Unsupervised reconstruction [26, 20, 13, 19].

2.1. 3D Supervised Reconstruction

3D supervised reconstruction directly trains a model to predict the 3D attributes, given many training images with ground truth 3D models (*e.g.* meshes, point cloud, or voxels). Pixel2mesh [39] and Mesh R-CNN [5] reconstruct mesh vertices as the shape attribute by iteratively sampling vertices' features and predicting their increment to ground-truth vertices. O-Net [25] predicts a 3D model shape attribute by classifying whether the randomly sampled 3D points are inside or outside the object. It achieves high 3D reconstruction performance and does not work if no 3D ground-truth is available. Our SMR does not need any 3D ground truth annotation and still performs 3D attribute-level supervision through our self-supervised methods.

2.2. 2D Supervised Reconstruction

CMR [15] is the first to reconstruct category-specific 3D models in the wild by 2D supervised reconstruction. It predicts 3D attributes from a single image, and utilizes the differentiable renderer [16] to re-project the reconstructed 3D model back to 2D image space. It finally adopts 2D supervised methods, including image/silhouette reconstruction and landmark regression [15], to train the network. Since shape and camera attributes are complicated to be separately encoded [18, 8], these 2D supervised methods usually require well-calibrated camera parameters possibly pre-calculated by SfM [29]. UMR [20] does not require camera parameters. But it still needs an external SCOPS [12] model to provide semantic parts as the prior information. In contrast, our SMR reconstructs 3D attributes without these additional prior models of camera calibration, category-specific template mesh, and semantic part model [12], which make it easier to be trained in an end-to-end manner.

2.3. Unsupervised Reconstruction

Recent work [41, 13, 19] further avoids 2D and 3D annotation by unsupervised learning. Unsup3d [41] achieves impressive reconstruction accuracy from only a collection of single images using the symmetric property. It only applies to symmetric and angle-limited objects (*e.g*., human faces). Cycle Consistency (CC) was proposed in CycleGAN [45] and was widely used in different unsupervised learning tasks [18, 19, 27, 21, 37]. CSM [18] and SSV [27] utilize geometric cycle consistency to predict pose parameters in an unsupervised manual, where the prediction network can correctly reproduce the pose of synthesis images. MUNIT [10] makes use of cycle consistency to disentangle images into content and appearance attributes. Navaneet *et al*. proposes shape cycle consistency for unsupervised point cloud reconstruction [19].

Our reconstruction supervision also inherits the property of cycle consistency. The difference is that ours is a unified framework to generate fine-grained distribution by interpolation, experimentally effective in predicting 3D mesh attributes.

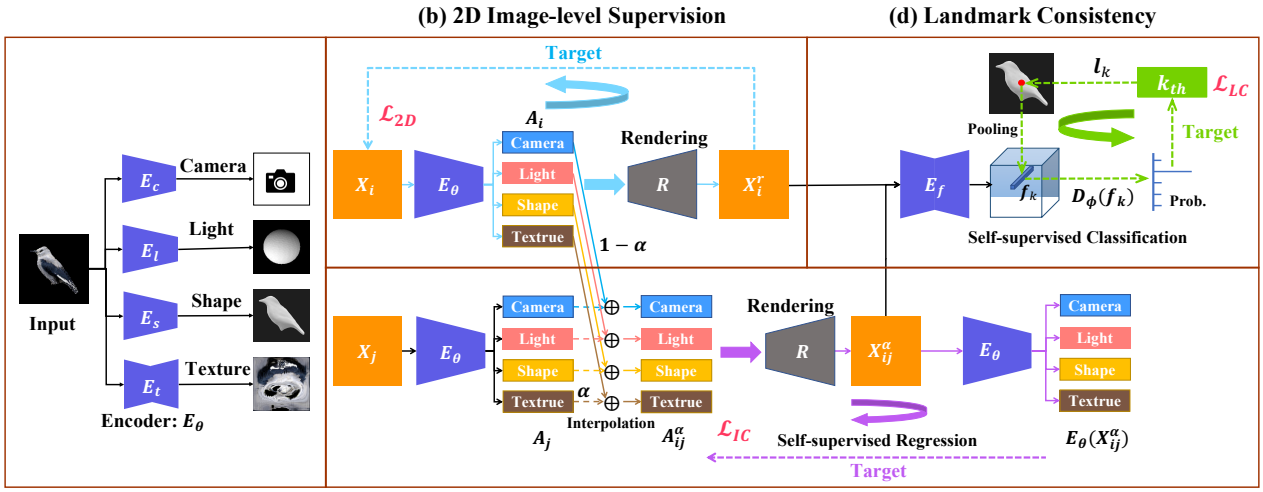

Figure 2: Overview of our Self-supervised 3D Mesh Reconstruction (SMR). Without any 3D ground-truth annotation, our Reconstructed Model (a) in Sec.3.2 can be trained to predict 3D mesh attributes from single images through 2D image-level supervision (b) in Sec.3.3.1, Interpolated consistency (c) in Sec.3.3.2, and Landmark Consistency (d) in Sec.3.3.3.

3. Approach

Given a collection of category-specific images with 2D silhouette annotation, we aim to train an encoder to reconstruct the camera, shape, texture, and light attributes of 3D mesh objects from single images.

3.1. Differentiable Rendering

To begin with, we briefly introduce the classic 3D mesh model and differentiable rendering. Let $O(S,T)$ denote a 3D mesh object. Shape attribute $S \in \mathbb{R}^{V\times 3}$ represents the mesh vertices (and faces). The total number of vertices is V. Texture attribute $T \in \mathbb{R}^{\mathcal{H}\times\mathcal{W}\times 3}$ represents the UV map with resolution $\mathcal{H}\times\mathcal{W}$. Let $C=(a,e,d)$ be the rendering camera, in which $a\in[0°,360°]$, $e\in[-90°,+90°]$, and $d\in(0,+\infty]$ stand for the the azimuth, elevation and distance parameters. Light attribute $L\in\mathbb{R}^l$ is modeled by Spherical Harmonics [33], which consists of a different spherical basis of angular frequency. l is the dimension of coefficient.

Given 3D attributes $A=[C,L,S,T]$, a 3D object $O(S,T)$ can be rendered as the 2D image and silhouette $X^r=[I^r,M^r]$ under camera view C and in lighting environment L. X^r is concatenated by the projected RGB image $I^r\in\mathbb{R}^{H\times W\times 3}$ and the silhouette mask $M^r\in\mathbb{R}^{H\times W\times 1}$ in channel axis. The rendering process 3D object is formulated as

$$X^r = R(A) = R([C,L,S,T]) \tag{1}$$

where R is a differentiable renderer, equivalent to a differentiable operation and does not contain any trainable parameters.

3.2. Reconstruction Model

For a category-specific dataset, the i_{th} input single image $I_i^r\in\mathbb{R}^{H\times W\times 3}$ and its silhouette $M_i^r\in\mathbb{R}^{H\times W\times 1}$ are concatenated as the input $X_i=[I_i,M_i],(i=1,2,...,N)$ in channel axis, where N is the number of training samples. 3D mesh reconstruction is to train an encoder E_θ that predicts 3D mesh attributes from a single input as

$$A_i=[C_i,L_i,S_i,T_i]=E_\theta(X_i), \tag{2}$$

where θ is the trainable parameter of the encoder. Note that E_θ is exactly the inverse process of R.

We independently predict attributes by four simple sub-encoders, as illustrated in Fig. 2(a). For camera encoder E_c, we predict a 4D vector consisting of $[a_x,a_y,e,d]$, where e and d represent the elevation and distance parameters of the camera. a_x and a_y denote the Cartesian coordinates of azimuth, and $a=atan2(a_x,a_y)$[1]. We calculate the azimuth parameter this way to avoid the discontinuous regression problem in the definition domain $[0°,360°]$.

The shape encoder E_s predicts the relative shape increment ΔS to a spherical mesh S_0, and $S=S_0+\Delta S$ calculates the object shape attribute. For texture encoder E_t, rather than directly outputting the texture UV map by an encoder-decoder model, we first predict a 2D flow map, and then apply spatial transformation [14] to generate texture UV map T. Similar strategies were also taken in CMR [15], which output texture with higher quality. Finally, for the light attribute, sub-encoder E_l directly encodes a l-dimension vector as the Spherical Harmonics model coefficient.

3D Supervision for Attribute Learning For the i_{th} image, if its 3D ground truth attributes A_i^{gt} are available, we

[1] https://en.wikipedia.org/wiki/Atan2

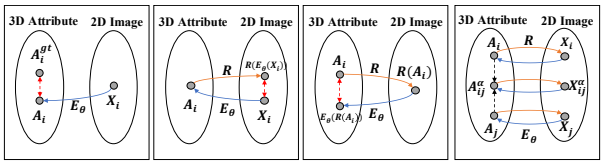

Figure 3: Different supervised reconstruction methods from 2D image space to 3D attribute space.

directly train the encoder to predict attribute $A_i = E_\theta(X_i)$ through regression as

$$\theta = \arg\min_\theta \frac{1}{N}\sum_{i=1}^{N} \parallel E_\theta(X_i) - A_i^{gt} \parallel_1 . \quad (3)$$

As illustrated in Fig. 3(a), we take the performance of using 3D supervision as the upper bound of general 3D object reconstruction.

3.3. Self-supervised Mesh Reconstruction

In this section, we describe how our method trains the encoder E_θ to learn 3D attributes at the 2D image level and the 3D attribute-level self-supervised learning.

3.3.1 2D Image-Level Supervision

As mentioned in Sec. 3.1, differentiable renderer R links 2D image space to 3D attribute space. Thus, we first optimize E_θ by 2D image-level supervision, formulated as

$$\theta = \arg\min_\theta \frac{1}{N}\sum_{i=1}^{N} Dist(R(E_\theta(X_i)), X_i), \quad (4)$$

where $Dist(\cdot,\cdot)$ indicates the distance between the reconstructed data $X_i^r = [I_i^r, M_i^r] = R(E_\theta(X_i))$ and input data $X_i = [I_i, M_i]$. This process is illustrated in Fig. 3(b). Similar to most 2D supervised reconstruction [3, 22, 20], we adopt image distance and silhouette distance loss to measure their difference.

Image Distance Foreground of the rendered and input images should be close by the L_1 distance. Thus we represent the image distance loss as

$$\mathcal{L}_{img} = \frac{1}{N}\sum_{i=1}^{N} \parallel I_i \odot M_i - I_i^r \odot M_i^r \parallel_1, \quad (5)$$

where $\odot$ denotes element-wise multiplication.

Silhouette Distance Besides, we utilize mask IoU loss to ensure that the projected silhouette M_i^r is identical to the ground truth silhouette M_i. Thus, the silhouette distance loss is written as

$$\mathcal{L}_{sil} = \frac{1}{N}\sum_{i=1}^{N}(1 - \frac{\parallel M_i \odot M_i^r \parallel_1}{\parallel M_i + M_i^r - M_i \odot M_i^r \parallel_1}). \quad (6)$$

Finally, the overall 2D image-level supervision is to minimize the weighted sum over above distance losses as

$$\mathcal{L}_{2D} = \lambda_{img}\mathcal{L}_{img} + \lambda_{sil}\mathcal{L}_{sil}, \quad (7)$$

where λ_{img} and λ_{sil} are the weights. $\mathcal{L}_{2D}$ supervises the reconstruction model by back propagating the loss gradient to the encoder E_θ through the differentiable renderer R.

3.3.2 Interpolated Consistency

2D image-level supervision in Section 3.3.1 can only optimize E_θ under original viewpoints of single images. Experimental results in Table 1 show that there is a large gap to 3D supervised reconstruction. To improve the reconstruction accuracy, we utilize the characteristics of category-specific 3D mesh model to perform 3D attribute-level supervision.

Our first motivation is to treat the encoded attributes A_i as the 3D annotations of the rendered image $R(A_i)$, and optimize E_θ through self-supervised regression. Therefore, apart from minimizing Eq. (4), we also optimize

$$\theta = \arg\min_\theta \frac{1}{N}\sum_{i=1}^{N} \parallel E_\theta(R(E_\theta(X_i))) - E_\theta(X_i) \parallel_1 . \quad (8)$$

Discussion There is representation of cycle consistency, as illustrated in Fig. 3(c). CSM [18] adopts this supervised method. The problem is that the number of encoded attributes is limited and might result in overfitting or degenerate reconstruction [19]. To alleviate the problem, random augmentation strategies were used to generate novel attributes in [27, 19, 30]. However, if we do not know the prior distribution of these 3D attributes, it might generate distorted body structure or out-of-view images, which affects the training process.

Our solution, differently, is to obtain the novel 3D annotations by linear interpolation, as illustrated in Fig. 3(d). For any pair of encoded attributes $A_i = [C_i, L_i, S_i, T_i]$ and $A_j = [C_j, L_j, S_j, T_j]$, we control the interpolation by a 4D vector $\alpha = [\alpha_c, \alpha_l, \alpha_s, \alpha_t]$ sampled in a uniform distribution $U \sim (0, 1)$ as

$$A_{ij}^\alpha = [C_{ij}^\alpha, L_{ij}^\alpha, S_{ij}^\alpha, T_{ij}^\alpha] = (1-\alpha)\cdot A_i + \alpha \cdot A_j. \quad (9)$$

The advantage is that it effectively generates a large number of fine-grained 3D mesh attributes, following similar distributions of original dataset, as illustrated in Fig. 4. We explain the physical meaning of each 3D attribute interpolation in the following.

Camera Interpolation The camera attribute C consists of azimuth, elevation, and distance parameters. Interpolation between different camera attributes $C_{ij}^\alpha = (1-\alpha_c)\cdot C_i + \alpha_c \cdot C_j$ can provide rendered image in their middle viewpoints. This can improve E_c's sensitivity and facilitate novel-view image synthesis.

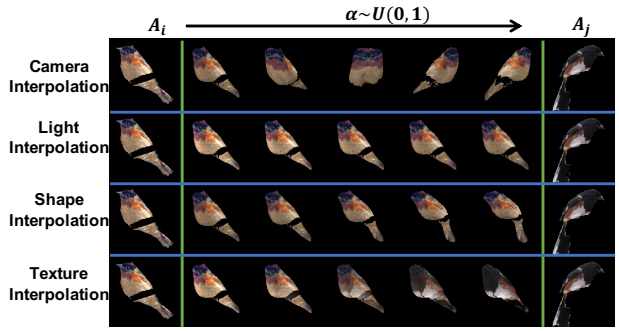

Figure 4: Rendered images with interpolated attributes. All interpolated attributes are treated as the 3D annotations to their render images. We train the reconstruction model through self-supervised leaning.

Light Interpolation The light attribute L is a l-d Spherical Harmonics coefficient. By interpolation $L_{ij}^{\alpha} = (1-\alpha_l) \cdot L_i + \alpha_l \cdot L_j$, we generate a variety of gradually changed lighting environment, enhancing the light sub-encoder and improving the quality of rendered images.

Shape and Texture Interpolation Each vertex of mesh of a category-specific object represents a specific landmark of object [15]. Thus we also interpolate shape and texture attributes by $S_{ij}^{\alpha} = (1-\alpha_s) \cdot S_i + \alpha_s \cdot S_j$ and $T_{ij}^{\alpha} = (1-\alpha_t) \cdot T_i + \alpha_t \cdot T_j$. This process can construct novel 3D models following the original geometrical and appearance distribution. Noted that texture attribute is represented by 2D maps, making texture interpolation similar to Mixup [44] in image augmentation.

These interpolated 3D attributes A_{ij}^{α} serve as the ground-truth 3D annotations to the rendered image X_{ij}^{α}, written as

$$X_{ij}^{\alpha} = R(A_{ij}^{\alpha}). \tag{10}$$

We then predict the 3D attributes of X_{ij}^{α} via encoder E_θ. IC loss $\mathcal{L}_{IC}$ is employed to train encoder E_θ by self-supervised regression as

$$\mathcal{L}_{IC} = \frac{1}{N} \sum_{i=1}^{N} \parallel E_\theta(X_{ij}^{\alpha}) - A_{ij}^{\alpha} \parallel_1 . \tag{11}$$

IC provides various fine-grained 3D models as the annotations to separately train each sub-encoder, which avoids the requirement of 3D ground-truth annotation when performing 3D supervision.

3.3.3 Landmark Consistency

Our proposed IC promotes the learning process of 3D attributes by introducing self-supervised 3D supervision. However, we notice that a few parts of reconstructed objects are still not realistic enough, as illustrated in Fig. 5. To further improve the local region quality of reconstructed objects, we propose landmark consistency as another self-supervised method. Our motivation is that feature representation of landmarks in all original and rendered images should be consistent. For instance, suppose the k_{th} landmark represents the center of left eye of a bird in one 3D model, in other 3D models, it should also have the same semantic meaning.

Specifically, as shown in Fig. 2(d), for the input image, we first extract its pixel-level feature maps F, like [36], by a U-Net [35] encoder E_f, and then project each mesh vertex to 2D image space as a landmark. Next, we calculate the landmark's location l_k and pool the local feature $f_k = F(l_k)$ by spatial transformation [14] from the feature maps. Finally, we adopt a Multi-Layer Perceptron (MLP) $D_\phi(\cdot)$ with weight ϕ to predict index category of each landmark. Since ground-truth category of f_k is also k, we build a self-supervised classification system for landmarks to train the 3D attribute encoder E_θ, as

$$\mathcal{L}_{LC} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{V} v_k y_k \log(D_\phi(f_k)), \tag{12}$$

where y_k is a one-hot vector with size V, in which only the k_{th} value is 1, and v_k indicates if k_{th} vertex is visible.

Our proposed LC maximizes feature distances of different landmarks and minimizes the distances of the same landmarks, which promote E_θ to reconstruct consistent and distinguishable landmarks for higher reconstruction quality in local regions.

3.3.4 Overall Loss

Finally, we combine supervised reconstruction at both image and attribute levels as the overall training loss for encoder E_θ as

$$\mathcal{L} = \lambda_{2D}\mathcal{L}_{2D} + \lambda_{IC}\mathcal{L}_{IC} + \lambda_{LC}\mathcal{L}_{LC}, \tag{13}$$

where λ_{2D}, λ_{IC}, and λ_{LC} respectively control the weights of 2D supervision, IC, and LC. During testing, we reconstruct the 3D mesh object from single images by predicting the 3D attributes through E_θ.

4. Experiments

To evaluate the effectiveness of our method, we first introduce the datasets and metrics in Sec. 4.1 and combine IC and LC with the 2D supervised and unsupervised reconstruction in Sec. 4.2. Sec. 4.3 compares SMR with start-of-the-arts. We then reconstruct more objects in the wild in Sec. 4.4. Finally, we show SMR's application in image synthesis in Sec. 4.5. Network structures and more experimental results are included in the supplementary material.

4.1. Datasets and Metrics

Datasets: We perform single-view 3D reconstruction experiments on the **ShapeNet** [2], **BFM** [41] and **CUB-200-2011** [38] datasets. ShapeNet is a large-scale synthesized

Encoder	Supervision	Airplane	Bench	Dresser	Car	Chair	Display	Lamp	Speaker	Rifle	Sofa	Table	Phone	Vessel	Mean
E_θ	3D	**46.4**	**32.5**	**53.6**	**57.8**	**39.9**	**42.0**	**49.7**	**57.9**	**43.4**	**46.3**	**38.8**	**50.6**	**54.7**	**47.2**
E_θ	2D	42.9	30.0	50.4	55.6	36.1	34.8	45.2	54.9	38.5	41.1	35.8	42.1	50.1	42.8
E_θ	2D + CC [18]	43.9	30.2	50.8	56.1	36.8	36.3	45.5	55.1	39.2	41.7	36.2	44.5	50.2	43.5
E_θ	2D + IC	44.7	31.1	52.8	56.9	39.1	39.2	48.2	57.4	42.6	44.8	37.6	49.3	53.8	46.0
E_θ	2D + IC + LC	**45.2**	**31.6**	**53.2**	**57.6**	**39.7**	**39.5**	**48.9**	**57.6**	**42.9**	**45.6**	**38.1**	**49.8**	**54.0**	**46.5**

Table 1: Comparison among different supervised methods for 3D Reconstruction on ShapeNet by shape 3D IoU.

Methods	Annotations					Mask IoU (%, ↑)	SSIM (%, ↑)	PCK (%, ↑)	FID (↓)
	Camera	Template	Landmarks	Parts [12]	Silhouette Mask				
CMR [15]	✓	✓	✓	✗	✓	73.8	44.6	28.5	115.1
CSM [15]	✗	✓	✗	✗	✓	-	-	48.0	-
DIB-R [3]	✓	✗	✗	✗	✓	75.7	-	-	-
UMR [20]	✗	✗	✗	✓	✓	73.4	71.3	58.2	83.6
SMR (Ours)	✗	✗	✗	✗	✓	**80.6**	**83.2**	**62.2**	**79.2**

Table 2: Comparison between our SMR and state-of-the-arts on the CUB-200-2011 dataset by multiple metrics. Ours only requires silhouette annotations and achieves better reconstruction in the original image view and novel view with higher Mask IoU, SSIM and PCK. It yields lower novel-view FID.

3D CAD dataset, containing 3D ground truth models of common object categories, *e.g.*, car, chair, and bench. We adopt the same train/test split provided by Soft-Ras [22] to evaluated the accuracy of 2D supervised reconstruction. BFM (Basel Face Model) [32] is a synthetic face prior model, and [41] built a 3D face reconstruction dataset based on it. We perform experiments on this dataset to evaluate the effect of IC and LC when combined with unsupervised reconstruction. CUB-200-2011 is a category-specific bird dataset consisting of single images and 2D annotations, such as 2D masks and landmarks. Recently, many methods [18, 15, 3, 20] evaluate their performance of single-view 3D reconstruction on this dataset and UMR [20] achieved state-of-the-art performance.

Evaluation Metrics: On the ShapeNet dataset, we evaluate the 3D reconstruction accuracy by 3D Intersection of Union (3D IoU) [22] between the reconstructed and ground truth 3D voxels of objects. On the BFM dataset, we measure the reconstructed depth metric Scale-Invariant Depth Error (SIDE) and Mean Angle Deviation (MAD) [41]. On the CUB-200-2011 dataset, although it does not contain any 3D annotation, we compare our model with state-of-the-art methods through the quality of synthesized images under the original and novel views.

The original view reconstruction is evaluated by Mask IoU and SSIM [40] between the reconstructed and input data. Since CUB-200-2011 has keypoint annotations, we also report Percent of Correct Keypoints (PCK) metric [18] that evaluates the accuracy of keypoint transfer for visible keypoints. PCK also indicates the performance of 3D object reconstruction. The novel-view reconstruction is evaluated by image generation metric FID [9]. We calculate the mean FID of the synthesized images in the novel view from 0° to 360° at an interval of 30°.

Implementation Details: The spherical mesh has $V = 642$ vertices and $1,280$ faces, same as those of [15, 20, 3]. The resolution of input images and texture UV maps is 256×256, except the shapenet dataset where the resolution is 64×64 [22]. The light parameter $l = 9$. The loss weights $\lambda_{img} = \lambda_{sil} = 10$, $\lambda_{2d} = \lambda_{IC} = 1.0$, and $\lambda_{LC} = 0.1$, which are obtained by grid search. We adopt DIB-R [3] as our differentiable renderer since it is applicable to all these 3D attributes. During training, the learning rate is initialized as 1×10^{-4} and decays 0.8 every 30 epoch. The optimizer is Adam [17] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$.

4.2. Comparison with Different Supervisions

Reconstruction on ShapeNet: Our SMR is a new supervised method designed for category-specific 3D mesh reconstruction. Therefore, our first experiment compares SMR with different supervised reconstruction baselines on the ShapeNet dataset, including 3D supervision, 2D supervision, and Cycle Consistency (CC) [18, 10, 19]. For fair comparison, all these methods adopt the same attribute encoder E_θ, as illustrated in Fig. 2(a). To compare the shape reconstruction accuracy on the ShapeNet dataset, a camera parameter is necessary to determine the scale and canonical viewpoint of the 3D object. There are a total of 13 categories on the ShapeNet, and we reconstruct them separately. The evaluation metric is shape's 3D IoU, and the results are shown in Table 1.

Analysis: For 2D supervised reconstruction, silhouette an-

Methods	SIDE ($\times 10^{-2}$, ↓)	MAD (deg. ↓)
3D Supervised	**0.410 ± 0.103**	**10.78 ± 1.01**
Average Depth	1.990 ± 0.556	23.26 ± 2.85
Unsup3D [41]	0.793 ± 0.140	16.51 ± 1.56
Unsup3D [41] + Random [30]	0.773 ± -	15.32 ± -
Unsup3D [41] + IC	0.762 ± 0.135	14.94 ± 0.135
Unsup3D [41] + LC	0.763 ± 0.139	14.64 ± 0.136
Unsup3D [41] + IC + LC	**0.758 ± 0.133**	**14.55 ± 0.131**

Table 3: Combination of our self-supervised methods with unsupervised 3D reconstruction on the BFM dataset.

IC				LC	PCK (%, ↑)	FID ↓
Camera	Shape	Texture	Light			
✗	✗	✗	✗	✗	42.6	174.1
✗	✓	✓	✓	✗	48.1	118.5
✓	✗	✓	✓	✗	52.8	101.3
✓	✓	✗	✓	✗	58.2	95.7
✓	✓	✓	✗	✗	58.9	92.6
✓	✓	✓	✓	✗	59.7	88.4
✓	✓	✓	✓	✓	**62.2**	**79.2**

Table 4: Effect of our proposed Self-Supervised Reconstruction modules.

notations are provided. Then 2D image-level supervision (in Sec. 3.3.1) is adopted to train the reconstruction model, which achieves 42.8% 3D IoU. For the 3D supervised reconstruction, since 3D attribute annotations are provided, we directly optimize the chamfer loss [34] between the predicted and ground truth shape attributes. 3D supervision obtained the highest 47.2% 3D IoU and surpassed 2D supervision by a large margin in all categories. It demonstrates the importance of 3D attribute-level supervision.

CC [18, 19] can be viewed as a baseline self-supervised method, which takes the *original* encoded attribute as the 3D annotation to the rendered image. Its α is randomly sampled to either 0 or 1, while our IC samples values between 0 and 1, as shown in Fig. 3(c)&(d). The experimental result shows that CC only improves accuracy by 0.7%. While ours achieves 46.5% 3D IoU when introducing IC and LC, significantly outperforms 2D and CC supervised methods and is even comparable with full 3D supervision. These experimental results validate that our method is useful to promote 2D supervised mesh reconstruction.

Reconstruction on BFM: In this experiment, we combine our proposed IC and LC with the unsupervised reconstruction methods [41, 30] on the BFM Face reconstruction dataset. We adopt the same network as Unsup3D and only introduce IC and LC for fair comparison. We evaluate the performance in terms of error of depth and normal reconstruction, *i.e*. SIDE and MAD [41]. The main difference between [30] and our IC is that the former randomly augments attributes. The results are shown in Table 3. 3D supervised reconstruction still achieves the best performance, while our method improves Unsup3D on both SIDE and MAD metrics, which manifest the effectiveness of IC and LC in 3D mesh reconstruction.

4.3. Comparison with State-of-the-arts

Quantitative Results Previous experiments were conducted on synthesized datasets with 3D annotations. In real world, most objects are photographed without 3D information. In this experiment, we compare our SMR with state-of-the-art methods [15, 3, 20] on CUB-200-2011 to demonstrate reconstruction performance for single images in the wild. The quantitative results are shown in Table 1. Compared with other methods that require cameras, landmarks, or parts annotations [12] annotations, ours are supervised by only silhouette annotation. Among these metrics, Mask IoU, SSIM, and PCK reflect the reconstructed accuracy under the original viewpoint, and we achieve significantly higher accuracy than others. FID reflects the mean reconstructed accuracy under a novel viewpoint from $0^\circ \sim 360^\circ$, and our SMR also achieves the best performance.

Ablation To validate the effect of each proposed consistency in our method, we perform an ablation experiment by removing one of them each time and test the reconstruction quality by the PCK of transferred keypoints and the FID of novel-view images as in Table 4. If there is only 2D supervision without any IC, both PCK and FID scores are the worst, explaining the importance of IC. Also, we note that the camera and shape IC are relatively more critical than texture and light IC, indicating that camera and shape attributes should be preferentially optimized. By introducing LC, our method reconstructs objects with the best performance, and the percent of correct key points is also significantly improved.

Qualitative Results The qualitative results are shown in Fig. 5. The reconstructed objects of CMR [15] look reasonable in shape and rough in texture. State-of-the-art method UMR [20] works better than CMR [15]. However, it still contains errors around the edge and overall color. For our method, if there is only 2D image-level supervision, the performance is not good. After introducing IC, the visualized results are competitive with UMR [20]. Finally, LC further improves the reconstructed quality in local parts, such as eyes and swings. The qualitative results demonstrate the effect of our IC and LC. We also render the reconstructed 3D model under different camera parameters to synthesize novel-view images, as illustrated in Fig. 6.

4.4. More Reconstruction Results in the Wild

To validate generalization of SMR, we implement it on more category-specific objects, as shown in Fig. 7. The cow, motorbike, and horse images are collected from LSUN [43] datasets, and the silhouette masks are detected by *detectron2* [42]. Our method does not require any category-specific template mesh or semantic parts [20] to reconstruct the 3D models. It is a general method for single images in the wild.

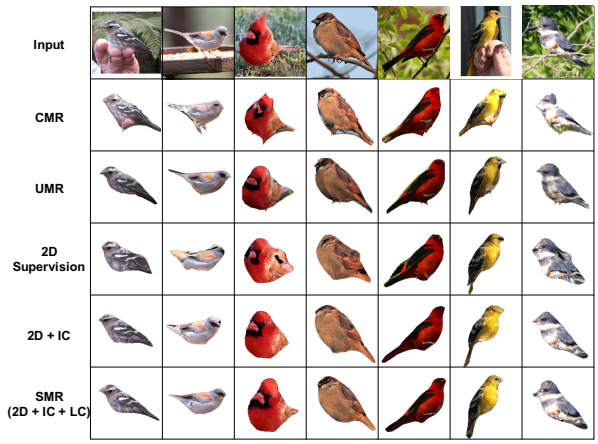

Figure 5: Qualitative comparison on the CUB-200-2011 dataset. Our SMR reconstructs object with more shape details and texture (best view by zoom-in).

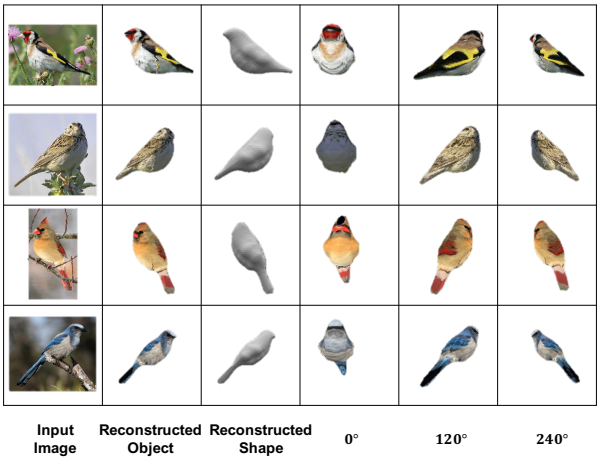

Figure 6: Novel-view object generation from single images. We encode the input image and modify the camera's azimuth parameter to render novel-view images.

4.5. Application in Image Synthesis

After obtaining the encoder E_θ, the above attribute interpolation (Eq. (9)) and rendering (Eq. (10)) process can synthesize images X_n from input X_a, X_b. We can control the camera, light, shape, and texture of X_n by setting different interpolation values $\alpha = [\alpha^c, \alpha^l, \alpha^s, \alpha^t]$ according to Eq. (9). We represent image synthesis as the following.

Camera Transfer: We can set $\alpha^c = 1$ and $\alpha^l = \alpha^s = \alpha^t = 0$ to perform camera transfer or novel-view synthesis as $X_n = R([C_b, L_a, S_a, T_a])$.

Shape Transfer: Similarly, setting $\alpha^s = 1, \alpha^c = \alpha^l = \alpha^t = 0$ changes the shape attribute to realize shape or pose transfer of $X_n = R([C_a, L_a, S_b, T_a])$.

Texture Transfer: Setting $\alpha^t = 1$ and $\alpha^c = \alpha^s = \alpha^l = 0$ means replacing the original texture UV map, which realizes texture transfer as $X_n = R([C_a, L_a, S_a, T_b])$.

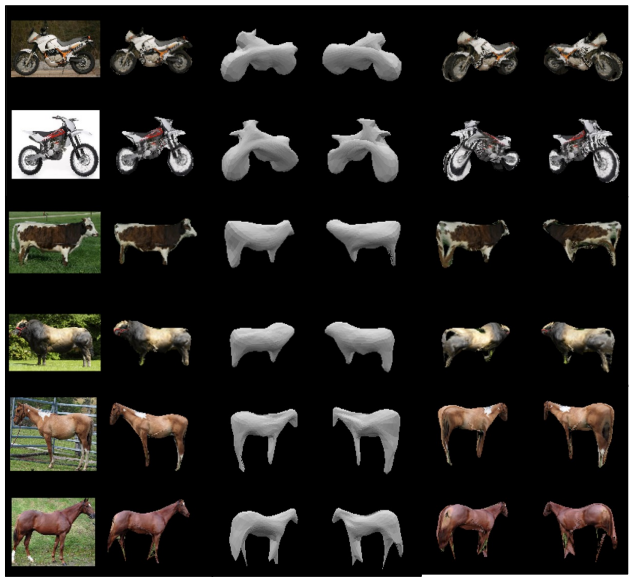

Figure 7: 3D object reconstruction in the wild by SMR.

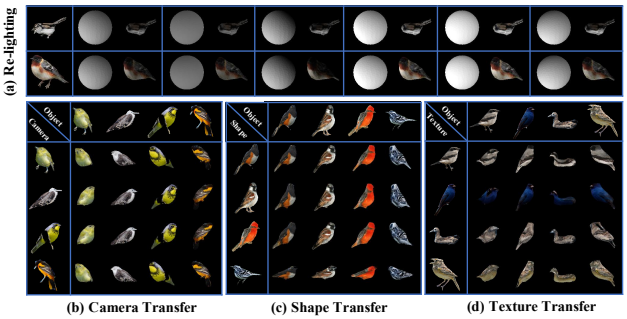

Figure 8: Different applications of SMR. We perform attribute transfer by replacing the original attribute.

Re-lighting: Since light attribute is also predict in our model, it is easy to perform image relighting by setting a different Spherical Harmonic coefficient L_r, as $X_n = R([C_a, L_r, S_a, T_a])$.

The visual results are shown in Fig. 8. Our SMR do not need additional networks for these image synthesis tasks.

5. Conclusion

We have proposed SMR, including 2D supervised, IC, and LC, to reconstruct 3D mesh from single images with only silhouette annotations. IC generates fine-grained 3D models to train the attribute encoder, and LC further improves the reconstruction quality in local regions. Our SMR improves both 2D supervised and unsupervised reconstruction and achieves state-of-the-art 3D reconstruction on multiple datasets. The main limitation of our method is the difficulty in modeling non-rigid objects, *e.g.*, human bodies. We leave it to future work for building a more general reconstruction method of deformed objects in the wild.

References

[1] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 1999. 1

[2] Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *CoRR*, 2015. 2, 5

[3] Wenzheng Chen, Huan Ling, Jun Gao, Edward J. Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. In *NeurIPS*, 2019. 1, 2, 4, 6, 7

[4] Haoqiang Fan, Hao Su, and Leonidas J. Guibas. A point set generation network for 3d object reconstruction from a single image. In *CVPR*, 2017. 1, 2

[5] Georgia Gkioxari, Justin Johnson, and Jitendra Malik. Mesh R-CNN. In *ICCV*, 2019. 1, 2

[6] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan C. Russell, and Mathieu Aubry. Atlasnet: A papier-mâché approach to learning 3d surface generation. *CoRR*, 2018. 1, 2

[7] Xian-Feng Han, Hamid Laga, and Mohammed Bennamoun. Image-based 3d object reconstruction: State-of-the-art and trends in the deep learning era. *CoRR*, 2019. 1

[8] Paul Henderson, Vagia Tsiminaki, and Christoph H. Lampert. Leveraging 2d data to learn textured 3d mesh generation. In *CVPR*, 2020. 1, 2

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 6

[10] Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 2, 6

[11] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: animatable reconstruction of clothed humans. In *CVPR*, 2020. 1, 2

[12] Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. SCOPS: self-supervised co-part segmentation. In *CVPR*, 2019. 2, 6, 7

[13] Eldar Insafutdinov and Alexey Dosovitskiy. Unsupervised learning of shape and pose with differentiable point clouds. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *NeurIPS*, 2018. 2

[14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NeurIPS*, 2015. 3, 5

[15] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 1, 2, 3, 5, 6, 7

[16] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *CVPR*, 2018. 1, 2

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[18] Nilesh Kulkarni, Shubham Tulsiani, and Abhinav Gupta. Canonical surface mapping via geometric cycle consistency. In *ICCV*, 2019. 2, 4, 6, 7

[19] Navaneet K. L., Ansu Mathew, Shashank Kashyap, Wei-Chih Hung, Varun Jampani, and R. Venkatesh Babu. From image collections to point clouds with self-supervised shape and pose networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020. 2, 4, 6, 7

[20] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Self-supervised single-view 3d reconstruction via semantic consistency. *CoRR*, 2020. 1, 2, 4, 6, 7

[21] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *ICCV*, 2019. 2

[22] Shichen Liu, Weikai Chen, Tianye Li, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *ICCV*, 2019. 1, 2, 4, 6

[23] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH Asia*, 2015. 1

[24] Matthew M. Loper and Michael J. Black. Opendr: An approximate differentiable renderer. In *ECCV*, 2014. 1

[25] Lars M. Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 1, 2

[26] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *CoRR*, 2020. 2

[27] Siva Karthik Mustikovela, Varun Jampani, Shalini De Mello, Sifei Liu, Umar Iqbal, Carsten Rother, and Jan Kautz. Self-supervised viewpoint learning from image collections. *CoRR*, 2020. 2, 4

[28] Michael Niemeyer, Lars M. Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*, 2020. 1, 2

[29] Penjani Nyimbili, Hande Demirel, Dursun Seker, and Turan Erden. Structure from motion (sfm) - approaches and applications. 2016. 2

[30] Xingang Pan, Bo Dai, Ziwei Liu, Chen Change Loy, and Ping Luo. Do 2d gans know 3d shape? unsupervised 3d shape reconstruction from 2d image gans. *CoRR*, 2020. 2, 4, 7

[31] Jeong Joon Park, Peter Florence, Julian Straub, Richard A. Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 1, 2

[32] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In Stefano Tubaro and Jean-Luc Dugelay, editors, *AVSS*, 2009. 6

[33] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *SIGGRAPH*, 2001. 3

[34] Eric Remy and Edouard Thiel. Computing 3d medial axis for chamfer distances. In *DGCI*, 2000. 7

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 5

[36] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Hao Li, and Angjoo Kanazawa. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, 2019. 5

[37] Kevin J. Shih, Aysegul Dundar, Animesh Garg, Robert Pottorf, Andrew Tao, and Bryan Catanzaro. Video interpolation and prediction with unsupervised landmarks. *CoRR*, 2019. 2

[38] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011. 2, 5

[39] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single RGB images. In *ECCV*, 2018. 1, 2

[40] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 6

[41] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *CVPR*, 2020. 1, 2, 5, 6, 7

[42] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 7

[43] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop. *CoRR*, 2015. 7

[44] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 5

[45] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2