Shape-Pose Ambiguity in Learning 3D Reconstruction from Images

Yunjie Wu,[1] **Zhengxing Sun,**[1] **Youcheng Song,**[1] **Yunhan Sun,**[1] **YiJie Zhong,**[1] **Jinlong Shi**[2]
[1] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, P R China
[2] Department of Computer, Jiangsu University of Science and Technology, Zhengjiang, P R China
szx@nju.edu.cn

Abstract

Learning single-image 3D reconstruction with only 2D images supervision is a promising research topic. The main challenge in image-supervised 3D reconstruction is the shape-pose ambiguity, which means a 2D supervision can be explained by an erroneous 3D shape from an erroneous pose. It will introduce high uncertainty and mislead the learning process. Existed works rely on multi-view images or pose-aware annotations to resolve the ambiguity. In this paper, we propose to resolve the ambiguity without extra pose-aware labels or annotations. Our training data is single-view images from the same object category. To overcome the shape-pose ambiguity, we introduce a pose-independent GAN to learn the category-specific shape manifold from the image collections. With the learned shape space, we resolve the shape-pose ambiguity in original images by training a pseudo pose regressor. Finally, we learn a reconstruction network with both the common re-projection loss and a pose-independent discrimination loss, making the results plausible from all views. Through experiments on synthetic and real image datasets, we demonstrate that our method can perform comparably to existing methods while not requiring any extra pose-aware annotations, making it more applicable and adaptable.

Introduction

Recovering 3D shape from a single image is a classical ill-posed problem in computer vision, requiring the prior of 3D shapes. Different from common 3D-supervised reconstruction methods (Choy et al. 2016; Fan, Su, and Guibas 2017; Wang et al. 2018; Mescheder et al. 2019), it becomes popular to learn an image-supervised 3D reconstruction recently. The core idea in image-supervised reconstruction is the *re-projection*, which requires the model to predict a shape that looks like the input image.

Many works use multiple views per object for training (Yan et al. 2016; Insafutdinov and Dosovitskiy 2018; Kato, Ushiku, and Harada 2018; Liu et al. 2019). However, learning to reconstruct with only single-view image collections remains challenging. The main challenge is the *shape-pose ambiguity*, which means an erroneous shape and pose can produce a 2D observation that closely matches the input image. This ambiguity can easily be eliminated with multi-view images because it is unlikely for an incorrect shape to match the multiple views at the same time. However, the ambiguity becomes difficult to resolve when we only have a single-view image for each object.

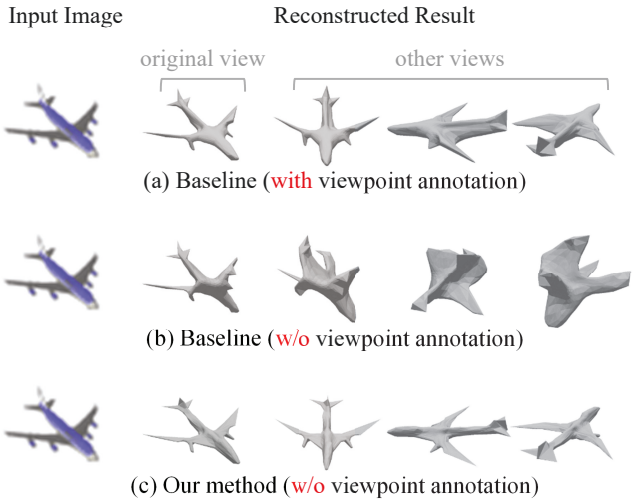

Figure 1: An example of shape-pose ambiguity. The three models are all trained with the same single-view image dataset. The two baselines are trained with the re-projection loss. (a) With the viewpoint annotations, the ambiguity is eliminated, and the model can learn meaningful reconstruction. (b) Without viewpoint annotations, the result looks plausible from the original viewpoint, but not correct from other viewpoints. (c) Our method can learn a correct 3D reconstruction without any viewpoint annotations.

Affected by the *shape-pose ambiguity*, we can not achieve meaningful reconstructions with the *re-projection* loss because the loss is pose-relevant and suffers from the ambiguity. An example is shown in Fig 1: if we provide the model with viewpoint annotations, then the pose is precise (the viewpoint and the pose are equivalent concepts in this work), and the ambiguity is eliminated. The model can learn

a plausible reconstruction (top row). If we block the viewpoint annotations and require the model to predict both the pose and the shape, it fails to produce meaningful results (middle row). The shape-pose ambiguity confuses the model and leads it to get inaccurate training signals from the re-projection. To handle the problem, some works rely on extra pose-aware annotation (Kato and Harada 2019; Kanazawa et al. 2018; Li et al. 2020) to help eliminate the ambiguity. However, these extra annotations are tedious and limit the adaptability. In this paper, we try to solve the problem without any extra pose-aware annotations.

The key to our method is a category-specific shape manifold estimation by a pose-independent GAN. We use a pose-independent discrimination loss instead of the pose-aware re-projection loss to help estimate the shape manifold. More specifically, a discriminator takes the training images from the same category of various poses as real samples and the projections of generated 3D shapes of any pose as fake samples. A 3D generator is required to fool the discriminator. As the discriminator is pose-independent, the generator has to produce plausible 3D shapes from any poses. Note that all images are from the same object category, so the discriminator can learn the effective category prior to guide the generator even without access to the pose annotations.

After learning the category-specific shapes manifold, we resolve the shape-pose ambiguity in the original images by training an image pose regressor. We train it with the rendered images of sampling shapes from the estimated shapes manifold. Then the pose can be disentangled from the images by predicting the pose label with the pose regressor. Finally, with the training images and the predicted poses, a reconstruction model is trained with both the re-projection loss from the predicted pose and a discrimination loss from other poses, making the results plausible from all views. Our contribution can be summarized as follows:

- We propose a 3D reconstruction framework which only requires unlabeled single-view image collections as training data.
- We reveal that, without extra pose-aware cues or annotations, it is possible to resolve the shape-pose ambiguity in single-view images from the same object category.
- We conduct comprehensive experiments on public datasets, including both synthetic and real images. The results show that our method can achieve comparable results to other methods using extra annotations.

Related Works

We overview related works from the aspects of supervision and classify them into 3D-supervised methods, multi-view-supervised methods, and single-view-supervised methods.

3D-supervised Methods

Many methods take the 3D annotations and corresponding 2D image as supervision. The representation of employed 3D annotation includes voxel (Choy et al. 2016; Wu et al. 2017), point cloud (Fan, Su, and Guibas 2017), mesh (Groueix et al. 2018; Wang et al. 2018), and sampling values on implicit 3D space function (Park et al. 2019; Mescheder et al. 2019; Chen and Zhang 2019; Michalkiewicz et al. 2019). The use of the 3D supervision provides full geometry and structure information of the target objects, which could be necessary for 3D reconstruction. With the development of the advanced 3D representation, well-designed network architecture, and effective loss functions, these methods achieve impressive results. However, the great difficulty in collecting full 3D supervision limits the scalability and adaptability of these methods.

Multi-view-supervised Methods

To relieve the burden of collecting 3D supervision, learning 3D reconstruction with 2D view supervision has recently become popular. Many view-supervised methods propose differentiable projection operation for comparing the 3D shapes and 2D observations (Yan et al. 2016; Insafutdinov and Dosovitskiy 2018; Kato, Ushiku, and Harada 2018; Liu et al. 2019). To handle the ambiguity in 2D observations, these works employ images from different viewpoints of the same object instance as training data. The multi-view images provide sufficient information to avoid the shape-pose ambiguity. The difference between projections of the generated 3D shape and ground truth images is computed as the loss to train the model. Although these methods avoid 3D supervision, multi-view images for each object instance are still challenging and expensive to obtain in practice.

Single-view-supervised Methods

Some works attempt to train a 3D reconstruction model with only single-view images. To resolve the shape-pose ambiguity, most of them rely on extra pose-aware annotations. Kato (Kato and Harada 2019) proposes to learn a 3D reconstructor by introducing adversarial learning to the unobserved views. They rely on the GT viewpoint labels for re-projection and discriminator's condition. Peng (Peng et al. 2021) proposes to disentangle the shape and the pose by introducing the domain adaptation, but they also rely on the GT viewpoint labels. Kanazawa (Kanazawa et al. 2018) attempts to predict both shape and pose matrix at the same time. The extra 2D keypoints are used to provide the information of pose transformation implicitly. Recently, Wu (Wu, Rupprecht, and Vedaldi 2020) proposes to learn the 3D reconstruction without pose-aware annotation. However, they only focus on symmetric objects. Besides, their 2D depth map representation is not suited to complex object categories, such as the chairs or tables. Differently, our method can learn to reconstruct a 3D mesh from complex categories.

Some other works do not use pose-aware annotations but they make some extra assumptions, including the existence of dense correspondence between each image and a 3D template (Li et al. 2020; Kulkarni, Gupta, and Tulsiani 2019; Kulkarni et al. 2020) or the uniform pose distribution (Henderson and Ferrari 2018). Differently, our work can recover the shape and pose information from single-view images without relying on these assumptions.

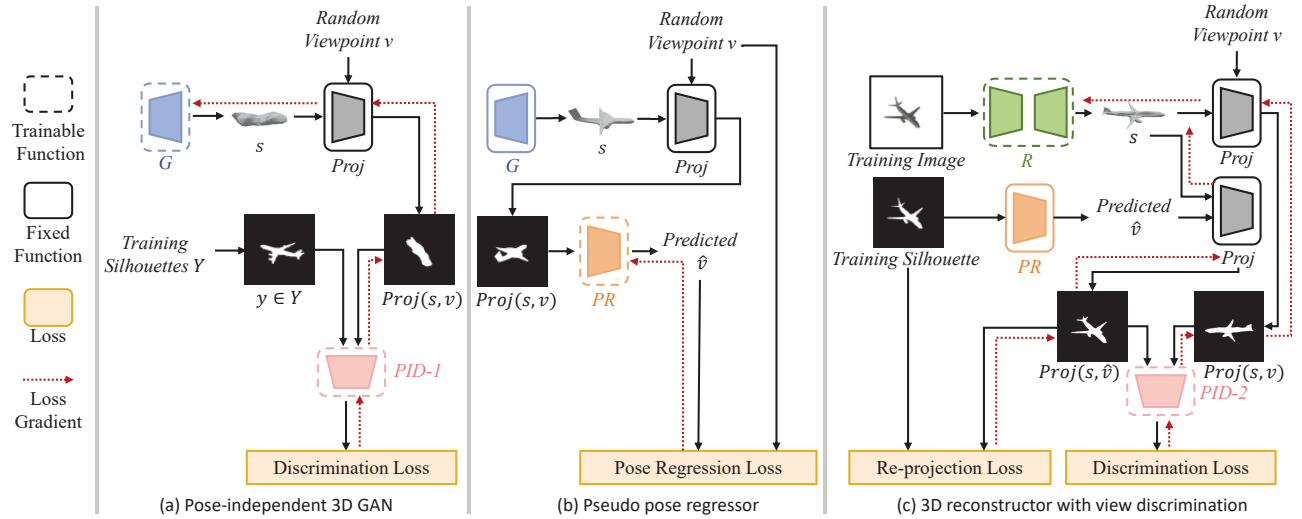

Figure 2: Overview of our method. It contains three main components: (a) a pose-independent 3D GAN for learning a category-specific shape manifold, (b) a pseudo pose regressor for predicting the pose from input silhouettes, (c) a 3D reconstruction network trained with both the re-projection and the view discrimination loss.

Proposed Method

Overview

Figure 2 shows an overview of our method. Our training data are single-view images I with silhouettes Y. Our method contains some modules: let $Proj(\cdot,\cdot)$ be a differentiable mesh renderer that takes a 3D mesh and a viewpoint and outputs the projection of the given shape from the given viewpoint, $G(\cdot)$ be a 3D generator that produces a 3D mesh from an input code, $D(\cdot)$ be a pose-independent discriminator that takes a 2D silhouette and outputs the probability that the input is from a target distribution, $R_{pose}(\cdot)$ be a pose regressor that predicts a silhouette's viewpoint, $R(\cdot)$ be a reconstructor that outputs a 3D mesh from the input RGB image. Then, in the first stage, we train the $G(\cdot)$ and the first $D(\cdot)$ to estimate a category-specific shape distribution. In the second stage, the $R_{pose}(\cdot)$ is learned with the sampling shapes from the estimated distribution. In the final stage, we employ the $R_{pose}(\cdot)$ to resolve the shape-pose ambiguity in the re-projection. Then we train the $R(\cdot)$ with the re-projection loss from the predicted pose and the view discrimination loss from other poses.

Parametrization for Shapes and Poses

The shape is parameterized as a deformable 3D mesh M. The initial shape is a unit sphere mesh, represented as vertices Ve and faces F. For a reconstructed shape M, the network predicts the vertices' transformation $\Delta Ve \in \mathbb{R}^{N*3}$, where N denotes the number of vertices. The shape is generated as $M = (Ve + \Delta Ve, F)$. We assume all reconstructed objects are up-oriented for the pose, and we do not consider the scale. With this setting, the shape's pose is equivalent to the camera's viewpoints. Following (Henderson and Ferrari 2018; Kato and Harada 2019), we model it as the viewpoint $v = (\theta, \phi)$, containing the azimuth $\theta \in [0, 2\pi)$ and a fixed elevation $\phi = 30$.

Pose-independent 3D GAN

As mentioned above, the main challenge in single-view supervised 3D reconstruction is the shape-pose ambiguity. Without extra pose-aware annotations, the model can hardly achieve the correct signal from the re-projection loss because it is related to the uncertain pose.

Unlike the pose-aware re-projection loss, we introduce a pose-independent discrimination loss to help a 3D generator estimate the category-specific shapes manifold. The $D_1(\cdot)$ is a silhouette discriminator, which is pose-independent and not concerned about the objects' poses. It takes training silhouettes as real samples and projections of generated 3D shapes as fake samples. The $G(\cdot)$ is a 3D generator, which produces a 3D shape from a noise code z and tries to fool the $D_1(\cdot)$. As $D_1(\cdot)$ does not distinguish images from different views, the $G(\cdot)$ has to produce the 3D shapes that look plausible from all views. In this way, we can estimate a manifold of plausible shapes with $G(\cdot)$ from the single-view images without viewpoints annotations. In practice, we employ the WGAN-GP (Gulrajani et al. 2017) to help stabilize the training. The loss for $D_1(\cdot)$ is:

$$L_{D_1}(Y, Z, V) = \sum_{z\sim Z, v\sim V} D_1(Proj(G(z), v)) - \sum_{y\sim Y} D_1(y) + \lambda \sum_{\hat{y}\sim\hat{Y}} \left(\|\nabla_{\hat{y}} D_1(\hat{y})\|_2 - 1\right)^2 \quad (1)$$

Where the Y denotes the training images' silhouettes. The Z denotes a Gaussian distribution for the sampling codes. The V denotes a uniform distribution of viewpoints. The $\hat{Y}$ denotes the interpolated space between real and fake samples (Gulrajani et al. 2017). The λ is a weighted parameter and is set to 10. The loss for $G(\cdot)$ is defined as:

$$L_G(Z,V) = -\sum_{z\sim Z, v\sim V} D_1(Proj(G(z),v)) \quad (2)$$

Pseudo Pose Regressor

After training the $G(\cdot)$, we have estimated the manifold of plausible 3D shapes. However, it is not enough to learn the reconstruction. As the shape-pose ambiguity still exists and we can not map the images to the shapes manifold.

To eliminate the ambiguity, we introduce a pseudo pose (viewpoint) regressor $R_{pose}(\cdot)$. Its input is 2D silhouettes and it predicts the viewpoint, as shown in Figure 2(b). The training data is generated by the trained $G(\cdot)$. We randomly generate 3D shapes with G and pick arbitrary viewpoints v. Then we project the generated shapes from the picked viewpoints. The $R_{pose}(\cdot)$ is optimized to predict the viewpoints. Note that the $G(\cdot)$ is free to determine its own canonical pose of the shape manifold. Hence the $R_{pose}(\cdot)$'s predicted pose coordinate is not aligned to the common canonical coordinate of the object category, so we call it *pseudo* pose regressor. The pose regression loss is defined as:

$$L_{R_{pose}}(Z,V) = \sum_{v\sim V, z\sim Z} Angle(v, R_{pose}(Proj(G(z),v))) \quad (3)$$

Where $Angle(\cdot,\cdot)$ means the loss between two viewpoints. Here we compute the angle between the two viewpoints. Especially, $\theta \in [0, 2\pi)$, the 0 and 2π correspond to the same azimuth, so we define the angle loss as:

$$\text{Angle}\ (v1, v2) = |\operatorname{atan2}(\sin(\theta 1 - \theta 2), \cos(\theta 1 - \theta 2))| \quad (4)$$

3D Reconstruction with View Discrimination

With the $R_{pose}(\cdot)$, we can easily resolve the shape-pose ambiguity in training images by predicting their viewpoint labels. Then a straightforward way to learn the reconstructor $R(\cdot)$ is training it with the re-projection loss:

$$L_R^{re}(I,Y) = \sum_i IoU\left(Proj\left(R\left(i_i\right), R_{pose}(y_i)\right), y_i\right) \quad (5)$$

Where the IoU denotes the intersection-over-union, the i_i, y_i denotes the i-th training image and silhouette. However, as shown in Figure 1(a), the results look worse from unobserved views when we train the model with only the re-projection loss. That is because, for the unobserved views, there is no direct supervision to guide the model.

A possible way to improve the unobserved views is to introduce the GAN to make these views plausible. In VPL (Kato and Harada 2019), they use a discriminator to distinguish observed and unobserved views. Their discriminator receives the pose annotations as extra conditions. This setting is not suited for our method. In our method, the $R_{pose}(\cdot)$'s predicted pose may not be strictly consistent between different object instances. Using the misaligned pose labels as conditions may disturb the discriminator from learning correctly across different objects. Hence, we replace the pose-aware discriminator with a pose-independent discriminator $D_2(\cdot)$. It aims to distinguish the projections from observed and unobserved views while it does not take extra pose labels as conditions. By confusing the observed and unobserved views, the supervision signal from the training images can help improve the unobserved views. The loss for the $D_2(\cdot)$ is defined as:

$$\begin{aligned} L_{D_2}(I,Y,V) = &\sum_{i,v\sim V, v\neq R_{pose}(y_i)} D_2((Proj(R(i_i),v))) \\ &- \sum_i D_2\left(Proj\left(R\left(i_i\right), R_{pose}\left(y_i\right)\right)\right) \\ &+ \lambda \sum_{\hat{p}\sim\hat{P}} \left(\left\|\nabla_{\hat{p}} D_2(\hat{p})\right\|_2 - 1\right)^2 \end{aligned} \quad (6)$$

Where $\hat{P}$ denotes the interpolated space between the projections from observed views and unobserved views.

To fool the $D_2(\cdot)$, the $R(\cdot)$ is required to pull the distribution of unobserved views' projections to the distribution of observed ones. The discrimination loss for $R(\cdot)$ is:

$$L_R^{dis}(I) = -\sum_{i,v\sim V, v\neq R_{pose}(y_i)} D_2((Proj(R(i_i),v))) \quad (7)$$

Following (Liu et al. 2019), we further impose a geometry loss L_R^{Geo} that regularizes the Laplacian of predicted shape to achieve appealing visual quality. The final loss of the $R(\cdot)$ is a weighted sum of the three losses:

$$L_R = L_R^{re} + \lambda L_R^{dis} + \mu L_R^{Geo} \quad (8)$$

In practice, the λ is set to 5e-4, and the μ is set to 5e-3 to make all loss terms in the close magnitude.

Experiments

Datasets

We test our method on two public datasets. The first is the ShapeNet's rendered images (Kato, Ushiku, and Harada 2018). Each object is rendered from 24 viewpoints. We use the provided train-test split. The second is CUB-200-2011 (Wah et al. 2011). It contains images of 200 species of birds. We crop the images around the birds. The open-wings birds are excluded due to the scarcity of them. We use an 8:2 train-test split. For both two datasets, we use no extra annotations except the silhouettes for training.

Baselines

We select some single-view supervised methods for comparison. For the ShapeNet, we set two trivial baselines. They employ the same network as ours and are trained with the re-projection loss. The $Baseline_s$ (B/L_s) is pose-supervised, having access to the GT viewpoints. The $Baseline_u$ (B/L_u) is pose-unsupervised, predicting the pose and the shape. We also compare with the state-of-the-art view priors learning (VPL) (Kato and Harada 2019). It is pose-supervised. For the CUB data, we select the state-of-the-art CMR (Kanazawa et al. 2018) for comparison.

Metric	Method	label-free	airplane	bench	dresser	car	chair	display	lamp	speaker	rifle	sofa	table	phone	vessel	Mean
IoU ↑	B/L_u	**Yes**	.345	.212	.301	.368	.252	.227	.214	.389	.230	.278	.218	.267	.296	.277
	B/L_s	No	.502	.362	.551	.560	.368	.381	.445	.586	463	.398	.428	.473	.445	.459
	VPL	No	.513	**.376**	**.591**	**.701**	**.444**	**.425**	.422	**.596**	**.479**	.500	**.436**	**.595**	**.485**	**.505**
	Ours	**Yes**	**.521**	.331	.565	.652	.439	.376	**.460**	.525	.468	**.574**	.397	.587	.460	.489
CD ↓	B/L_u	**Yes**	3.51	6.54	6.63	6.60	8.09	8.53	9.02	8.96	8.42	8.73	9.53	9.60	9.45	7.97
	B/L_s	No	2.28	3.69	2.84	2.15	6.40	4.26	9.40	5.47	4.04	6.14	4.74	3.91	2.85	4.47
	VPL	No	1.56	**2.39**	2.97	1.15	**3.29**	3.64	**4.51**	4.30	1.30	3.56	**3.80**	1.54	**1.67**	**2.74**
	Ours	**Yes**	**0.70**	2.80	**1.66**	**1.02**	4.90	**1.88**	11.87	**3.06**	**0.46**	**1.80**	9.77	**0.74**	3.24	3.38

Table 1: Quantitative comparison with the B/L_u, the B/L_s, and the VPL (Kato and Harada 2019). The metrics are IoU and CD(x0.01). Our method outperforms the baselines and is comparable to the VPL, while our method requires no pose labels.

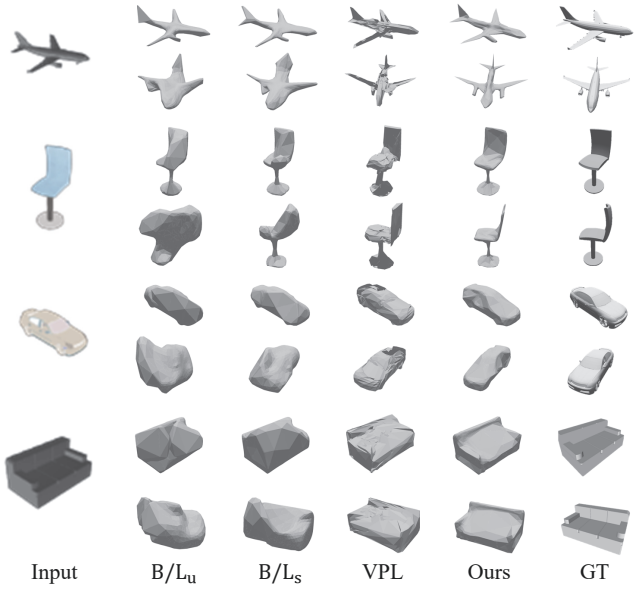

Figure 3: Qualitative comparison. We show the reconstructed shapes of each method from the input image's pose and another unobserved pose.

Metrics

We use the 3D intersection-over-union (IoU) and the chamfer distance (CD) for evaluation. The IoU is computed in 32 voxel resolution, and the chamfer distance is computed on 2048 surface sampled points. As there is no GT 3D data for the CUB dataset, all quantitative evaluations are conducted on the ShapeNet dataset. In particular, the canonical pose of the viewpoint-unsupervised methods (B/L_u and ours) are not consistent with the GT 3D data. So we search the rotation angle to align the results before evaluating them.

Implementation Details

Each shape has 642 vertices, the same as VPL (Kato and Harada 2019). We use the differentiable renderer (Liu et al. 2019) as $Proj(\cdot)$. The $G(\cdot)$ has three fully-connected layers,

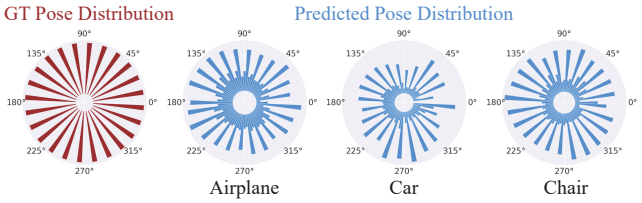

Figure 4: Distribution of the predicted poses.

generating a 642*3 transformation vector from a noisy vector. The $D(\cdot)$ has five 2D convolutional layers with ReLU activation. The $R_{pose}(\cdot)$ has five 2D convolutional and two fully-connected layers, with ReLU activation in between and a sigmoid activation in the final. The $R(\cdot)$ has a decoder the same as $G(\cdot)$ and an image encoder of resnet-18 (He et al. 2016). The batch size is 32, and the learning rate is 1e-4. We provide more details in the appendix and the code in: https://github.com/JiejiangWu/Shape-Pose-Ambiguity.

Results on Synthetic Image Dataset

Comparisons with Baselines We report the quantitative results in Table 1. The results suggest that our method outperforms the two trivial baselines. Improving over the B/L_s (in which the GT labels eliminates the shape-pose ambiguity) confirms that our method can resolve the shape-pose ambiguity in the single-view 3D reconstruction. Compared to the VPL, our method achieves comparable performance. Given their methods relies on the GT pose labels for training while our method does not require any pose-aware annotations, our method is more applicable and scalable.

We also present some qualitative reconstruction results in Figure 3. More results can be found in the supplementary material. It is clear that: 1) The B/L_u's results are plausible from the input poses but wrong from other poses, and the B/L_s's results are better from these poses. This confirms the influence of the shape-pose ambiguity. 2) Our results look plausible from all poses, indicating our method resolves the ambiguity successfully. 3) Our results are similar to VPL's in general outlines but more smooth on the surface. It is caused by the difference in the renderer module (Liu et al. 2019).

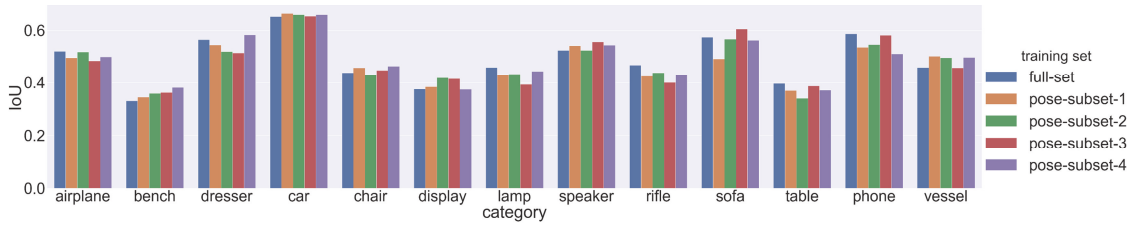

Figure 5: Performance of models trained on subsets. With a sparse and biased distribution of poses, the performance does not show noticeable degradation, illustrating our method is robust to the distribution of the training data's poses.

	airplane	car	chair
shape manifold	.476	.556	.361
learned reconstruction	.521	.652	.439

Table 2: Quantitative comparison on IoU between the shape manifold and the learned reconstruction.

This explains why our method is slightly worse than VPL on IoU while better on CD in most categories. Because the CD is highly sensitive to the surface, but the IoU is not.

Pose Prediction Our method can predict the pose of objects in an unsupervised way. However, as the model is free to determine the canonical pose, it is impossible to evaluate the result by comparing it with the GT pose labels. Instead, we visualize the distribution of both predicted and GT poses for qualitative evaluation. Especially, in the dataset, the elevation is fixed to 30, so we only visualize the azimuths' distribution. The Figure 4 shows the result. Our method's predicted distribution is close to the real distribution, demonstrating it has learned correct pose information.

Robust to the Distribution of Training Poses The poses in the dataset (Kato, Ushiku, and Harada 2018) are uniformly and densely distributed. However, in most image datasets, the poses are often distributed in a non-uniform and sparse way. To check whether our method is robust to the distribution of the training images' poses, we sample some subsets from the original dataset and train the model. Especially, let $0-23$ denotes the 24 poses (azimuths) in the dataset, we make four subsets with different poses distribution. First: $\{0, 2, 4, 6, 8, 12, 14, 16, 18, 20, 22\}$. It is relatively sparse in distribution. Second: $\{0, 6, 12, 18\}$. It contains four poses, which are extremely sparse. Third: $\{0, 4, 5, 6, 12, 18\}$. It is sparse and non-uniform distributed. Fourth: $\{0, 6, 10, 11, 12, 18\}$. It is similar to the third. We respectively train the model with four subsets and evaluate it with the same test set, containing images from all poses. The results are reported in Figure 5. We can find that, even trained with sparse and non-uniform poses, the model achieves similar performance. This experiment confirms our method's robustness.

Ablation Study Due to the page limitation, we only report three categories' results here. Others are in the appendix.

First, is the learned reconstruction necessary? With the learned shape manifold, an alternative way for reconstruction is to cooperate the $G(\cdot)$ with an image encoder, instead of training the $R(\cdot)$ from scratch. We report the quantitative comparison in Table 2. We can find that training the reconstructor from scratch significantly outperforms using the trained $G(\cdot)$. Some qualitative examples are shown in Figure 6. The shapes generated by $G(\cdot)$ miss some details (in the red box) but are improved and corrected (in the green box) with the learned reconstruction. The reason is that the $G(\cdot)$ is trained without the re-projection loss, so it can produce plausible shapes but probably ignore meaningful details. This experiment indicates, for 2D-supervised 3D reconstruction, optimizing the shape decoder with the re-projection loss is necessary and essential.

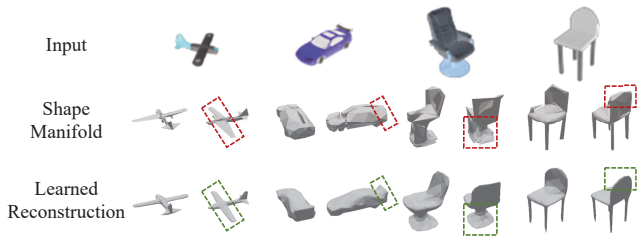

Figure 6: Qualitative comparison between the shape manifold and the learned reconstruction.

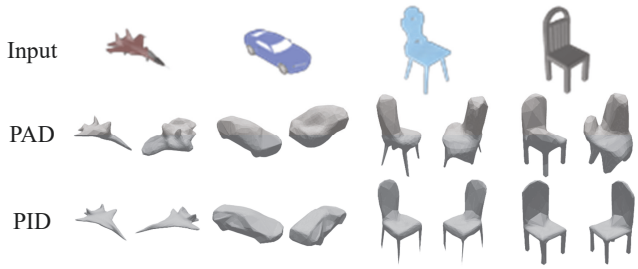

Figure 7: Qualitative comparison between the models trained with the PAD and the PID.

Second, is the pose-independent discriminator necessary? As mentioned above, we emphasized the importance of employing a pose-independent discriminator (PID) rather than a pose-aware discriminator (PAD). We validate this statement with an experiment. We replace the pose-independent discriminator $D_2(\cdot)$ with a pose-aware discrim-

	airplane	car	chair
PAD	.410	.344	.325
PID	.521	.652	.439

Table 3: Quantitative comparison on IoU between the models trained with the PAD and the PID.

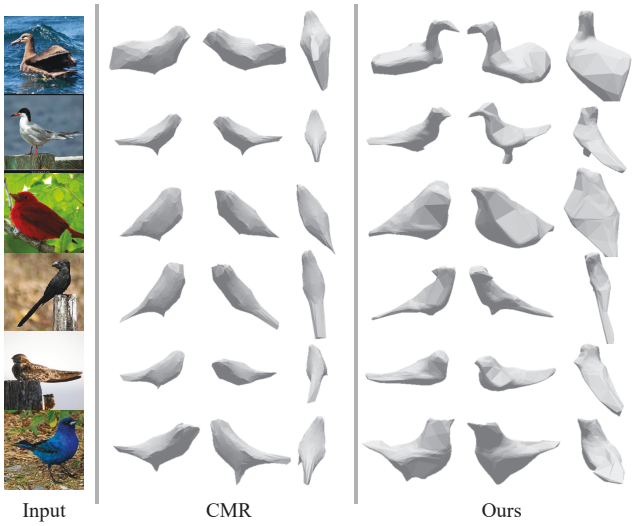

Figure 8: Qualitative comparison between the CMR (Kanazawa et al. 2018) and our method.

inator $PAD(\cdot)$, which receives the predicted viewpoint of the image from $R_{pose}(\cdot)$ as extra condition. Its architecture follows (Kato and Harada 2019). The quantitative results are reported in Table 3, and some qualitative results are shown in Figure 7. Obviously, the model with PID outperforms the model with PAD, verifying the necessity of using the pose-independent discriminator in our method.

Results on Real Image Dataset

Qualitative Reconstruction Results We test our method on the real image dataset: CUB-200-2011 (Wah et al. 2011). We compare our method with the CMR (Kanazawa et al. 2018), which uses extra 2D keypoints on this dataset for reconstruction. We present the qualitative results in Figure 8. It can be observed that 1) our results are more diverse in geometry and more similar to the input images. A possible reason is that CMR performs the SFM on 2D keypoints to obtain a template mesh. This process may restrict their method of generating diverse results. 2) CMR's results are more plausible than ours in some details. For example, the abdomen of the bird in row 2,3. That is not surprising as they have access to the GT 2D keypoints, providing more information about the detailed geometry. Without using the 2D keypoints annotations, our method achieves comparable results to CMR.

Images Clustering by Pose With our method, we can learn the pose information contained in the dataset. We use the pose regressor to predict a viewpoint label for each image and visualize the pose distribution in Figure 9. As there

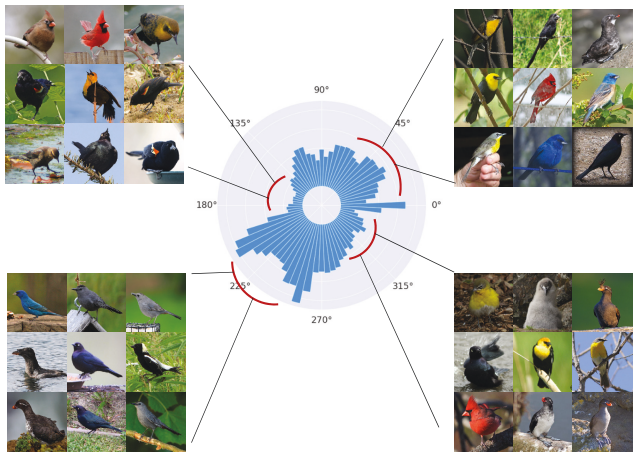

Figure 9: The distribution of predicted poses in CUB. We visualize the distribution and some sampled images.

is no ground-truth pose for comparison, we additionally visualize some sampled images for some cluster bins. From the visualization, we can tell that: 1) The images in CUB-200-2011 are not balanced in the pose. The images are taken more from the sides than from the front. 2) The images from the same bin share similar poses, confirming that our method has learned correctly to resolve the shape-pose ambiguity, even for the natural images with various appearances.

Conclusion and Future Work

In this paper, we propose to learn the 3D reconstruction by resolving the shape-pose ambiguity in single-view images. Experimental results on both synthetic and natural images show that 1) it is possible to resolve the shape-pose ambiguity in images from the same object category, without any pose-aware labels or annotations. 2) We can achieve comparable results to previous 3D reconstruction methods, demonstrating that the pose-aware annotations are unnecessary in 2D-supervised 3D reconstruction.

Though our method improves performance on image-supervised 3D reconstruction, it relies on silhouettes for training. Training end-to-end 3D reconstruction and silhouette segmentation would be a promising future direction. Besides, another interesting direction is to introduce the implicit 3D representation(Mescheder et al. 2019; Park et al. 2019), which may further improve the results.

Acknowledgments

This work was supported by National High Technology Research and Development Program of China (No. 2007AA01Z334), National Natural Science Foundation of China (Nos. 61321491 and 61272219),National Key Research and Development Program of China(Nos. 2018YFC0309100, 2018YFC0309104), the China Postdoctoral Science Foundation (Grant No. 2017M621700) and Innovation Fund of State Key Laboratory for Novel Software Technology (Nos. ZZKT2018A09).

References

Chen, Z.; and Zhang, H. 2019. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5939–5948.

Choy, C. B.; Xu, D.; Gwak, J.; Chen, K.; and Savarese, S. 2016. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, 628–644. Springer.

Fan, H.; Su, H.; and Guibas, L. J. 2017. 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*, 605–613.

Groueix, T.; Fisher, M.; Kim, V.; Russell, B.; and Aubry, M. 2018. AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In *CVPR 2018*.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved training of wasserstein gans. In *Advances in neural information processing systems*, 5767–5777.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Henderson, P.; and Ferrari, V. 2018. Learning to Generate and Reconstruct 3D Meshes with only 2D Supervision. In *British Machine Vision Conference (BMVC)*.

Insafutdinov, E.; and Dosovitskiy, A. 2018. Unsupervised learning of shape and pose with differentiable point clouds. In *Advances in Neural Information Processing Systems*, 2802–2812.

Kanazawa, A.; Tulsiani, S.; Efros, A. A.; and Malik, J. 2018. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 371–386.

Kato, H.; and Harada, T. 2019. Learning view priors for single-view 3d reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9778–9787.

Kato, H.; Ushiku, Y.; and Harada, T. 2018. Neural 3d mesh renderer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3907–3916.

Kulkarni, N.; Gupta, A.; Fouhey, D.; and Tulsiani, S. 2020. Articulation-aware Canonical Surface Mapping. In *Computer Vision and Pattern Recognition (CVPR)*.

Kulkarni, N.; Gupta, A.; and Tulsiani, S. 2019. Canonical Surface Mapping via Geometric Cycle Consistency. In *International Conference on Computer Vision*.

Li, X.; Liu, S.; Kim, K.; De Mello, S.; Jampani, V.; Yang, M.-H.; and Kautz, J. 2020. Self-supervised Single-view 3D Reconstruction via Semantic Consistency. In *ECCV*.

Liu, S.; Li, T.; Chen, W.; and Li, H. 2019. Soft Rasterizer: A Differentiable Renderer for Image-based 3D Reasoning. *The IEEE International Conference on Computer Vision* .

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

Michalkiewicz, M.; Pontes, J. K.; Jack, D.; Baktashmotlagh, M.; and Eriksson, A. 2019. Deep Level Sets: Implicit Surface Representations for 3D Shape Inference. *arXiv preprint arXiv:1901.06802* .

Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; and Lovegrove, S. 2019. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 165–174.

Peng, B.; Wang, W.; Dong, J.; and Tan, T. 2021. Learning Pose-invariant 3D Object Reconstruction from Single-view Images. *Neurocomputing* 423: 407–418. ISSN 0925-2312.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001. URL http://www.vision.caltech.edu/visipedia/CUB-200-2011.html.

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

Wu, J.; Wang, Y.; Xue, T.; Sun, X.; Freeman, B.; and Tenenbaum, J. 2017. Marrnet: 3d shape reconstruction via 2.5 d sketches. In *Advances in neural information processing systems*, 540–550.

Wu, S.; Rupprecht, C.; and Vedaldi, A. 2020. Unsupervised Learning of Probably Symmetric Deformable 3D Objects from Images in the Wild. In *CVPR*.

Yan, X.; Yang, J.; Yumer, E.; Guo, Y.; and Lee, H. 2016. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *Advances in Neural Information Processing Systems*, 1696–1704.