Conditional Single-view Shape Generation for Multi-view Stereo Reconstruction

Yi Wei[1,2*]   Shaohui Liu[1,2*]   Wang Zhao[1,2*]   Jiwen Lu[1†]

[1]Department of Automation, Tsinghua University, Beijing, China
[2]Department of Electronic Engineering, Tsinghua University, Beijing, China

`b1ueber2y@gmail.com`, {`wei-y15, zhaowang15`}`@mails.tsinghua.edu.cn`, `lujiwen@tsinghua.edu.cn`

Abstract

In this paper, we present a new perspective towards image-based shape generation. Most existing deep learning based shape reconstruction methods employ a single-view deterministic model which is sometimes insufficient to determine a single groundtruth shape because the back part is occluded. In this work, we first introduce a conditional generative network to model the uncertainty for single-view reconstruction. Then, we formulate the task of multi-view reconstruction as taking the intersection of the predicted shape spaces on each single image. We design new differentiable guidance including the front constraint, the diversity constraint, and the consistency loss to enable effective single-view conditional generation and multi-view synthesis. Experimental results and ablation studies show that our proposed approach outperforms state-of-the-art methods on 3D reconstruction test error and demonstrates its generalization ability on real world data.

1. Introduction

Developing generative models for image-based three-dimensional (3D) reconstruction has been a fundamental task in the community of computer vision and graphics. 3D generative models have various applications on robotics, human-computer interaction and autonomous driving, etc. Researchers have discovered effective pipelines on reconstructing scene structures [11, 32] and object shapes [15, 41]. Recently, inspired by the promising progress of deep learning on 2D image understanding and generation, much great work has been done on using differentiable structure to learn either volumetric or point cloud predictions from single-view [6, 39, 47] and multi-view [4, 14] images.

Despite the rapid progress on the task of single-view image-based shape reconstruction, there remains a fundamental question: *can a single image provide sufficient information for three-dimensional shape generation?* It is in-

---

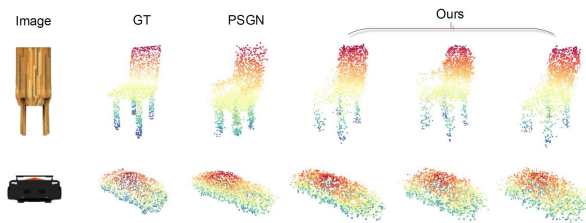*indicates equal contribution
†corresponding author

Figure 1: In real-world scenarios, one single image cannot sufficiently infer a single 3D shape due to occlusion. While our predictions can handle the ambiguity including the chair arm and the car length, deterministic methods can only predict one mean shape which is not necessarily correct. We further extend this for multi-view stereo reconstruction.

tuitive that in one picture took or rendered from a specific view, only the front of the object can be seen. Ideally, most existing methods implicitly assume that the reconstructed object has a relatively symmetrical structure, which enables reasonable guess on the back part. However, this assumption may not be true when it becomes much more complex in real world scenarios.

In this paper, we address the problem of modeling the uncertainty for single-view object reconstruction. Unlike conventional generative methods which reconstruct shapes in a deterministic manner, we propose to learn a conditional generative model with a random input vector. As the groundtruth shape is only a single sample of the reasonable shape space for a single-view image, we use the groundtruth in a partially supervised manner, where we design a differentiable front constraint to guide the prediction of the generative model. In addition, we use a diversity constraint to get the conditional model to span the space more effectively. Conditioning on multiple random input vectors, our conditional model can give multiple plausible shape predictions from a single image.

Furthermore, we propose a synthesis pipeline to transfer the single-view conditional model onto the task of multi-view shape generation. Different from most existing methods which utilize a recurrent unit to ensemble multi-view features, we consider multi-view reconstruction as taking

the intersection of the predicted shape space on each single-view image. By introducing a simple paired distance metric to constrain the multi-view consistency, we perform online optimization with respect to the multiple input vectors in each individual conditional model. Finally, we concatenate the multi-view point cloud results to obtain the final predictions.

Our training pipeline benefits from pre-rendered depth image and the camera pose without explicit 3D supervision. By modeling the uncertainty in single-view reconstruction via a partially supervised architecture, our model achieves state-of-the-art 3D reconstruction test error on ShapeNet-Core [3] dataset. Detailed ablation studies are performed to show the effectiveness of our proposed pipeline. Additional experiments demonstrate that our generative approach has promising generalization ability on real world images.

2. Related Work

Conditional Generative Models: Generative models conditioned on additional inputs are drawing continuous attention with its large variety of applications. Conditional Generative Adversarial Networks (CGAN) [25] made use of the concept of adversarial learning and yielded promising results on various tasks including image-to-image translation [12] and natural image descriptions [5]. Another popular trend is the variational autoencoder (VAE) [17]. Conditional VAEs achieved great success on dialog generation [34, 50]. Different from the CGAN [25] scenario, we only have limited groundtruth observations in the target space, which relates to the concept of one-shot learning [19]. In this paper, we propose a partially supervised method with a diversity constraint to help learn the generative model. Then, we introduce a synthesis method on multiple conditional generative models.

Deep Single-view Reconstruction: With the recent advent of large 3D CAD model repositories [3, 22, 36, 46], large efforts have been made on deep single-image reconstruction in 3D vision. While conventional methods [4, 15, 44] focused on volumetric generation, point cloud and mesh representation were used in recent literature [6, 18, 23, 8]. Researchers have introduced various single-view reconstruction approaches including 2.5D sketches [43, 49, 45], adversarial learning [2, 44], generating novel views [23, 28, 35, 37], re-projection consistency [10, 39, 40, 47, 51], high resolution generation [13, 38] and structure prediction [21, 26]. Some recent post-processing attempts include point cloud upsampling [48] and shape inpainting [42]. These methods except [6] all implicitly assumed that with prior knowledge the network can fantasy the missing part in the input image. However, in real world scenarios, the back part of the object may be much too complex to infer. This was recently addressed by [45]. In this

work, we propose to model the ambiguity for the task of single-view point cloud reconstruction. Different from [6] which used a relatively simple MoN loss to enable multiple predictions, we focus on different treatments between the front part and the back part of the object and improve its representation ability.

Deep Multi-view Synthesis: Multiple images took or rendered from different views contain pose-aware information towards 3D model understanding. Conventional methods [52] utilized RGB-D cameras for 3D reconstruction via estimated correspondence [24, 30, 16]. For RGB-based deep multi-view reconstruction, most existing methods [4, 14] utilized a recurrent unit to integrate the extracted features from each single view. [23, 37] used concatenation to get dense point cloud predictions. For a specific CAD model, the reconstruction results from different views should be consistent. This consistency was used [10, 39, 40, 47] as a supervisory signal via re-projection for unsupervised single-view generative model training. In this work, we introduce a multi-view synthesis techniques by online optimizing the multi-view consistency loss with respect to the random inputs on conditional models.

3. Approach

3.1. Overview

The problem of single-view shape reconstruction was conventionally formulated as a one-to-one mapping $\phi : I \rightarrow S$, where I denotes the input RGB image and S denotes the predicted shape. This one-to-one generative model was widely used to output either voxels [4] or point clouds [6] via cross entropy loss and differentiable distance metrics. Most existing methods took the implicit assumption that the input image is sufficient to predict the whole shape structure.

Consider the probabilistic model $p(S|I)$, where S is a random shape conditioned on the input image I. In perfect conditions where useful knowledge is completely learned from the groundtruth shape, the existing deterministic architecture $\phi : I \rightarrow S$ can learn the most probable shape S^*, where

$$\phi(I) = S^* = E_S[p(S|I)] \tag{1}$$

Most existing single-view reconstruction methods utilized this deterministic formulation and could generalize relatively well to the test set. This is probably due to the fact that most objects in the widely used ShapeNet [3] dataset have a symmetric or category-specific structure which enables reasonable inference. However, this is arguably not true especially in complex scenarios. In fact, the structure of the occluded back part is usually relatively ambiguous.
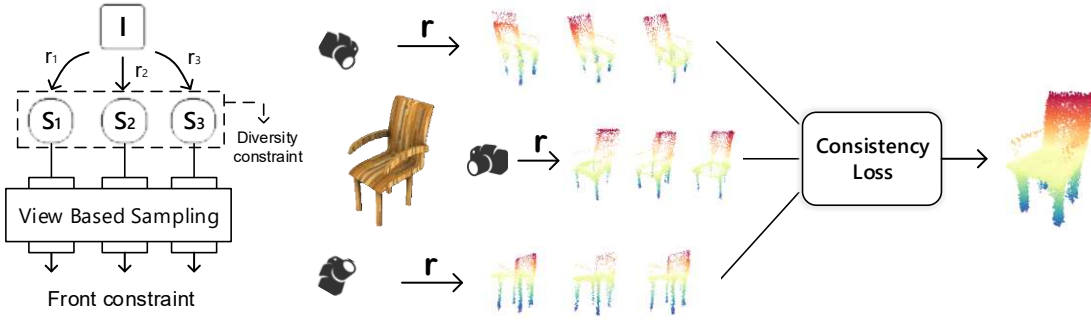
Figure 2: Overview of our proposed appoach. **Left:** the single-view training pipeline of the model. One single image is fed with a set of random inputs $\{r_i\}_{i=1}^n$ to get $S_i = f(I, r_i)$. Then, the partially supervised front constraint is used along with a diversity constraint to enable the model to focus more on the front part while maintaining generating diversity. **Right:** Inference. With different random inputs $\{r_i\}_{i=1}^n$, our conditional generative model $G : S = f(I, r|\theta)$ can generate multiple plausible shapes from each view. The consistency loss is used to synthesize the multiple conditional generative model to get the final predictions.

To better model this inherent ambiguity, we introduce a conditional generative model $f : I \times r \to S$, where the image-based generation is conditioned on a Gaussian input vector r. We aim to learn a mapping to approximate the probabilistic model $p(S|I)$ in the reasonable shape space.

However, different from the scenarios of generation in CGAN [25], we only have limited groundtruth (in fact, only one shape per image) which cannot span the reasonable shape space. Motivated by the fact that the front part can be sufficiently inferred from the input image while the back part is relatively ambiguous, we propose to train the conditional generative model in a partially supervised manner. Furthermore, we introduce a diversity constraint to help span the reasonable space.

Let us take a step further into a multi-view scenario, where the problem comes to a many-to-one mapping. Considering only three input views for simplicity here, we have $I_1 \times I_2 \times I_3 \to S$. As most parts of the object are covered by different input views, we assume that multi-view reconstruction can be viewed as a deterministic inference with sufficient information. Given a conditional generative model $S = f(I, r)$, the output S^* follow the constraint:

$$S^* = f(I_1, r_1) = f(I_2, r_2) = ... = f(I_n, r_n) \quad (2)$$

From the Bayesian perspective, Single-view and multi-view reconstruction can be formulated as to approximate $p(S|I_i)$ and $p(S|\{I_i\}_{i=1}^n)$. In this paper, the idea is to first implicitly approximate $p(S|I_i)$ with a conditional model and then apply it to deterministic multi-view synthesis, which differs from conventional RNN-based methods [4, 14]. This choice has two main reasons. 1) The data scale is limited and only one 3D groundtruth exists for every image, making it not easy to explicitly parameterize $p(S|I_i)$.

2) For a 3D shape, rendered images from different views is correlated. Thus, it is relatively intractable to formulate $p(S|\{I_i\}_{i=1}^n)$ with $p(S|I_i)$ and directly optimize maximum likelihood. Same problem exists in many other research directions eg. multi-view pose estimation. We propose an alternate conceptual idea to get intersections of manifolds conditioned on different views with assistance of pair-wise distance minimization. Figure 2 summarizes an overview of our proposed approach.

3.2. Modeling the Uncertainty for Single-view Reconstruction

Given a specific architecture on a conditional generative model $S = f(I, r)$, we aim to learn the ambiguity of single-view reconstruction from limited groundtruth data. In this section, we first briefly review differentiable distance metric in the shape space, then introduce two of our proposed differentiable constraints to help learn the conditional model.

Distance Metric: Two existing differentiable distance metrics between point sets were originally used in [6]. These metrics are Chamfer Distance (CD) and Earth Mover's distance (EMD) [31]. CD finds the nearest neighbor and is formulated as below[3] in Eq.(3), while EMD learns a optimal transport between two point sets in Eq.(4).

$$d_{CD}(S_1, S_2) = \sum_{x \in S_1} \min_{y \in S_2} \|x - y\|_2 + \sum_{x \in S_2} \min_{y \in S_1} \|x - y\|_2 \quad (3)$$

$$d_{EMD}(S_1, S_2) = \min_{\phi : S_1 \to S_2} \sum_{x \in S_1} \|x - \phi(x)\|_2^2 \quad (4)$$

---

[3]We use the first-order version of Chamfer Distance following [23].

We use both of these two metrics in our training pipeline. Following [23], the two terms in Chamfer Distance were jointly reported as pred→GT and GT→pred at test stage.

Front Constraint: We propose a *front constraint* along with a new differentiable operation: *view based sampling*, which enables the conditional model to learn in a partially supervised manner. Different from recently proposed point cloud downsampling strategies [20, 29] which aims to uncover inner relationship for coarse-to-fine understanding, our proposed *view based sampling* layer outputs a set of points which consist the front part of the shape from a specific view.

The overview of the *front constraint* is shown in Figure 3. In the proposed approach, we get the generative model to focus more on the front N_1 points, while the $N - N_1$ remaining points are conditionally generated. For the *view based sampling* layer, we first render the point cloud onto a 2D depth map with the intrinsic and extrinsic parameters given. Then, we sample all of the points which contribute to the rendered map. This strategy enables that all sampled points are on the front side of the object from the view. Note that because pixel-wise loss on the depth map used in [23] is only differentiable on the rendered z axis, it will not work in our single-view training scenario (See Section 4.4).

By either applying view based sampling to the groundtruth point cloud or applying inverse-projection to the pre-rendered depth map, we can get the groundtruth front part. Then, CD or EMD [31] can be used to acquire the loss of the front constraint $loss_{front}$ and differentiably guide the sampled point cloud.

Diversity Constraint: Because for one input image, only one groundtruth shape is available at the training stage, simply training the conditional generative model with groundtruth constraints will hardly get the model to span the reasonable shape space. For different input vectors r, we aim to get different predictions which all satisfy the front constraint. With the hinge loss as in the widely used Triplet Loss [33] in face verification, we propose a *diversity constraint* which uses the Euclidean distance of input r as the distance margin in 3D space.

Specifically, considering paired input vectors r_1, r_2 for a single training image I, we have two predicted point clouds $S_1 = f(I, r_1)$ and $S_2 = f(I, r_2)$. The loss of the *diversity constraint* is formulated as in Eq.(5) below.

$$loss_{div} = \max(0, \|r_1 - r_2\|_2 - \alpha EMD(S_1, S_2)) \quad (5)$$

Because the counts of both point clouds are equal, we use EMD [31] to measure the distance between S_1 and S_2. The hyper-parameter α helps control the diversity of the predicted point clouds.
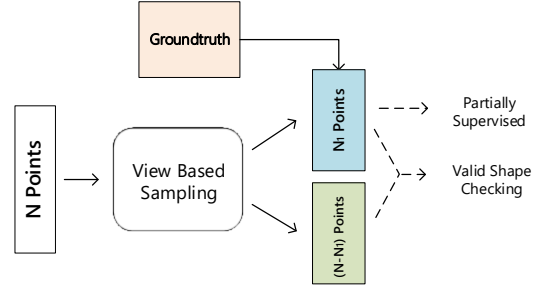

Figure 3: Partial supervision on the front part of the conditionally generated point clouds with *view based sampling*. The validity (whether the outputs form a valid shape) is ensured by a GAN checking module in Eq.(6).

Latent Space Discriminator: Combining the front constraint and the diversity constraint forms a initial paradigm for modeling the ambiguity of the single-view reconstruction. However, this paradigm puts little pressure on the back part. Thus, motivated by the generative adversarial networks [7] and recently proposed representation learning method [1] on point clouds, we propose to add a latent space discriminator to better learn the shape priors. Specifically, we first train an auto-encoder on the point cloud domain. Then, we transfer the decoder to the end of our architecture and get it fixed. Finally, we apply WGAN-GP [9] on the top of the latent space. Take E_S as the encoder from the point cloud domain to the latent space, and E_I as the encoder we use from the input image I_i and the random noise r_i to the latent variable z_i, the loss is formulated as below, where S is the sampled point cloud from the dataset.

$$
\begin{aligned}
loss_{gan} = &-\mathbb{E}_{I_i \sim p_{data}, r_i \sim p(r)}[D(E_I(I_i, r_i))] \\
&+ \mathbb{E}_{S \sim p_{data}}[D(E_S(S))] \\
&- \lambda \mathbb{E}_{\hat{z} \sim p_{\hat{z}}}[(\|\nabla_{\hat{z}} D(\hat{z})\|_2 - 1)^2]
\end{aligned} \quad (6)
$$

Training on Single-view Images: As discussed, we can train a conditional generative model using single input image with the optimization objective in Eq.(7) at the training stage. β, γ denotes the relative loss weight of the diversity loss and the GAN loss respectively.

$$loss = loss_{front} + \beta loss_{div} + \gamma loss_{gan} \quad (7)$$

The training is performed in an iterative min-max manner as the widely-used GAN training strategy. The hyper-parameter α and β modulates how far the generative model goes beyond the observed groundtruth.

3.3. Synthesizing Multi-view Predictions

Finetuning on Multi-view Images: To get the network to learn more clues on the high level structure of the object,

we finetune the single-view pretrained model on multi-view conditions. For synthesizing the multi-view point clouds at the training stage, we simply concatenate predicted point clouds from different views. Then, $loss_{front}$ was computed in different views on the concatenated results and $loss_{div}$ was computed in different random inputs to guide the training process. Specifically, for each shape, 8 views and 5 random inputs for each is used to train the model. Similar to the single-view training stage, we use the combined $loss$ in Eq.(7) as our minimization objective.

Inference: As shown in Eq.(2), from the deterministic perspective the multi-view reconstruction can be viewed as taking the intersection of the reasonable shape space conditioned on each input image. Thus, we propose a consistency constraint directly on the shape level. Consider a set of results $\{S_i\}_{i=1}^n$ from n different views, where $S_i = f(I_i, r_i)$. The consistency loss is formulated in Eq.(8).

$$loss_{consis} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} CD(S_i, S_j) \quad (8)$$

Figure 4 shows our inference method. The method of freezing the inference model and adjusting the input is popular in the field of adversarial attacks [7]. By online minimizing $loss_{consis}$ with respect to the input vectors $\{r_i\}_{i=1}^n$ in the conditional model, we get more consistent results. To prevent the optimization from local minimum, we use heuristic search in the $\{r_i\}_{i=1}^n$ initialization. **Algorithm 1** shows our detailed inference pipeline. Our method does not require camera calibration at inference.

4. Experiments

4.1. Experimental Settings

Network Architecture: Figure 5 briefly shows our network architecture. For the encoder-decoder branch, we used the two-branch version of the point set generation network in [6]. We set random input r as a 128-dimensional vector. The embedding branch employs a structure with two fully connected layers and two convolutional layers. Channel-wise concatenation is performed on the embedded vector z_r and the encoded features z_i. For more details on the two-branch network in [6], refer to our supplementary material.

Implementation Details: We trained our conditional generative network for two stages on a GTX 1080 GPU. The input images were rendered from ShapeNetCore.v1 [3] with the toolkit provided by [39]. To cover the entire object, we uniformly sampled the rendered views along the horizontal circle with a random longitudinal perturbation. We took 80% of the data for training and the rest for testing. We used $\lambda = 10$ and $\gamma = 0.1$ for adversarial learning. At the

---

Algorithm 1 Inference pipeline for multi-view reconstruction.

Input: multi-view (n views) images $\{I_i\}_{i=1}^n$, conditional generative model $G : S = f(I, r; \theta)$.
Output: predicted shape S.
  1: Randomly sample 5 groups of $\{r_{ij}\}_{j=1}^5$, each of which consists of n random inputs.
  2: Feedforward with $S_i = f(I_i, r_{ij}; \theta)$ compute the $loss_{consis}$ in Eq.(2) for each group. Denote the group with the minimum consistency loss $\{r_i^+\}_{i=1}^n$.
  3: Freeze the parameter θ of the inference model. Initialize $r_i = r_i^+$.
  4: Iteratively minimize $loss_{consis}$ until convergence, get the optimized inputs $\{r_i^*\}_{i=1}^n$.
  5: Feedforward with $S_i = f(I_i, r_i^*; \theta)$ and concatenate S_i to get the final prediction S.
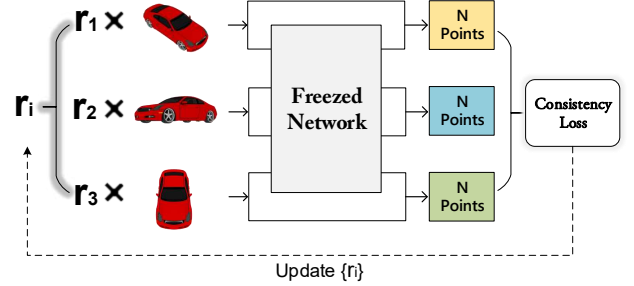

Figure 4: Multi-view inference by online minimizing the consistency loss.
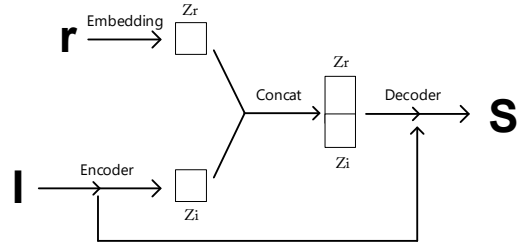

Figure 5: Brief overview of the network architecture.

first training stage, we trained the model using single-view images for 40,000 iterations with a batch size 16 and 5 random inputs for each image. $\alpha_1 = 0.2$, $\beta_1 = 10.0$. Then, we finetuned our model for 100,000 iterations on multi-view images. There were 2 shapes in each batch, 8 views for each shape, and 5 random inputs for each view. $\alpha_2 = 0.1$, $\beta_2 = 1.0$. We used Adam with an initial learning rate 1e-4 in both stages. At test stage, we used 8 views to reconstruct

Figure 6: Visualization of multiple predictions on a single image conditioned on random sampled r.

Table 1: Evaluation on the diversity of the conditional generative models.

Method	Consistency loss
EMD + MoN [6]	1.65
$loss_{front}$	0.55
$loss_{front}$ + MoN	2.52
$loss, \beta = 1.0$	2.88
$loss, \beta = 5.0$	3.18
$loss, \beta = 10.0$	**3.36**

the point clouds. The range of the longitudinal perturbation is a degree of $[-20, 40]$. Following [23], we scaled the reconstruction error CD by a factor of 100. **Code will be made available.**

4.2. Multiple Predictions on a Single Image

Our generative model is able to predict multiple plausible shapes conditioned on the random input r. As discussed in Section 3.2, the front constraint guides the generation of the front part, and the diversity constraint enables the conditional model to span the shape space.

Qualitative Visualization: Figure 6 visualizes the multiple predictions on a single input image conditioned on randomly sampled r. It is shown that our conditional model generates plausible shapes with a large diversity. The front part from the view of the input RGB image is predicted in a relatively more deterministic manner while the back part is mainly controlled by the random input r.

Evaluation on Uncertainty Modeling: We conducted experiments to better verify the generating diversity of the proposed conditional generative model. We took one single image I and randomly sampled 10 inputs $\{r_i\}_{i=1}^{10}$. Then, we fed the model with $S_i = f(I, r_i)$ and computed $loss_{consis}$ in Eq.(2) on the predicted shape set $\{S_i\}_{i=1}^{10}$. We re-implemented the fully-supervised MoN method in [6]. For fair comparison, we used the conditional model after the first training stage in this experiment. Table 1 shows the results. The partial supervision boosts the diversity of the predicted shapes. Moreover, it is demonstrated that when the loss weight β of the diversity loss rises, the generating diversity gets consistent increase.

Table 2: CD (FPS-CD) results of single-category experiments on ShapeNet [3] dataset. We compare our methods with existing methods including [4, 6, 23, 47].

Method	GT $\to$ pred	pred $\to$ GT	CD (FPS-CD)
3D-R2N2	2.47	3.21	5.68
PTN	1.86	2.60	4.46
PSGN	2.06 (2.06)	2.27 (**2.27**)	4.34 (4.34)
Lin *et al.*	1.66 (2.16)	2.35 (2.59)	4.01 (4.75)
Ours	**1.39** (**1.73**)	**1.98** (2.35)	**3.37** (**4.08**)

4.3. Multi-view Shape Reconstruction

Evaluation Metric: The most widely used metric on evaluating point cloud generation is the Chamfer Distance in Eq.(3). For comparison, we use the same protocol with [23]. However, it is worth noting that CD computation under different numbered point clouds is relatively confusing. Thus, we also report FPS-CD where we used farthest point sampling [29] to get same-numbered point clouds.

Single-category Experiments: In this experiment, we applied our conditional generative model on the task of single-category multi-view shape reconstruction on ShapeNet [3] "chairs". We re-implemented several widely used image-based reconstruction methods including 3D-R2N2 (5 views) [4], PTN [47], PSGN [6] and Lin *et al.* [23] on our synthetic dataset. We converted the voxels predicted by [4, 47] to point clouds in the experiment. For the groundtruth point clouds, we used the uniformly sampled point clouds directly from [1]. Note that our idea is also complementary to voxel-based deterministic methods (eg. MarrNet [43]), where metrics can be developed on voxel space and back-propagation of cross-entropy loss is performed only from the front. Here we use PSGN [6] with the point cloud outputs for direct comparison. Table 2 shows the experimental results. It is reported that although our conditional generative method is not only partially supervised but also without explicit 3D supervision at training stage, our approach outperforms all of the baseline methods.

Multi-category Experiments: We tested our model in multi-category experiments following [4] on 13 popular categories on ShapeNet [3] dataset. As shown in Table 3, our proposed method outperforms two baseline methods 3D-R2N2 [4] and PSGN [6] by a relatively large margin.

Qualitative Results: For qualitative analysis, in Figure 7 we visualize the predicted shapes for two state-of-the-art baseline methods: 3D-R2N2 [4] and PSGN [6]. It is shown that our partially supervised conditional generative model
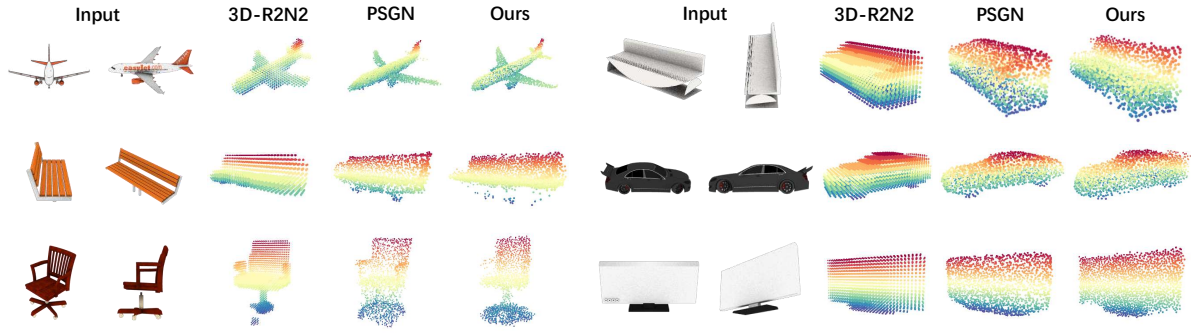
Figure 7: Qualitative comparison between ours and baseline approaches [4, 6].

Table 3: CD (FPS-CD) results of multi-category experiments on ShapeNet [3] dataset.

Category	3D-R2N2 [4]	PSGN [6]	Ours
airplane	5.25	2.89 (**2.89**)	**2.65** (3.10)
bench	5.39	4.30 (4.30)	**3.48** (**4.17**)
cabinet	4.60	4.87 (**4.87**)	**4.10** (5.39)
car	4.51	3.68 (**3.68**)	**3.06** (4.03)
chair	5.78	4.67 (4.67)	**3.80** (**4.64**)
display	5.69	5.96 (5.96)	**4.44** (**5.27**)
lamp	10.54	6.04 (**6.04**)	**5.15** (6.27)
loudspeaker	6.54	6.42 (6.42)	**4.99** (**6.39**)
rifle	4.38	3.22 (3.22)	**2.60** (**3.05**)
sofa	5.43	4.93 (**4.93**)	**4.31** (5.35)
table	5.31	4.45 (**4.45**)	**3.43** (4.51)
telephone	5.06	4.34 (**4.34**)	**3.50** (4.35)
watercraft	5.38	4.66 (4.66)	**3.57** (**4.24**)
all	5.68	4.39 (4.39)	**3.58** (**4.34**)

Table 4: Comparison between the conditional model and the deterministic model. Both CD and FPS-CD are reported.

Method	CD	FPS-CD
deterministic	3.62	4.18
conditional	**3.37**	**4.08**

can infer reasonable shapes which are dense and accurate. More details are generated due to the specific aim on the front parts of the objects.

4.4. Ablation Studies

Conditional vs. Deterministic: To demonstrate the effectiveness of the conditional model, we implemented a deterministic model $S = f_d(I)$. For fair comparison, we used an encoder-decoder structure similar with our network and trained the deterministic model for two stages with the front constraint. Single-category experiment was conducted on the deterministic model. Table 4 shows the results. Al-

Table 5: Ablation studies on the diversity constraint and the consistency loss. s1 denotes the pretraining on singleview images, while s2 denotes the finetuning process on multi-view images. s1 is always trained with diversity loss. The lossdiv in the table denotes the diversity loss specifically in s2. Experiments were conducted on the single-category setting. Both CD and FPS-CD is reported.

s1	s2	$loss_{div}$	$loss_{consis}$	CD	FPS-CD
✓	✗	✗	✗	3.97	6.30
✓	✗	✗	✓	3.87	5.77
✓	✓	✗	✗	3.51	4.18
✓	✓	✓	✗	3.40	4.16
✗	✓	✓	✓	3.52	4.24
✓	✓	✓	✓	**3.37**	**4.08**

though the shape in ShapeNet [3] dataset often has symmetric structure, the conditional generative model outperforms the deterministic counterpart by 0.25 on CD.

Analysis on different features in the framework: We performed ablation analysis on three different features: two-stage training, diversity constraint at multi-view training stage and consistency loss during inference. As shown in Table 5, all features achieve consistent gain on the final performance.

Front constraint vs. Projection loss: Our conditional model can be trained on single-view images with the front constraint and the diversity constraints. For comparison, we directly applied the projection loss used on multi-view images training in [23] on single-view images, the training did not converge. Because the pixel-wise loss on the depth map suffers from non-differentiable quantization in the render process, the projection loss can only get gradients from the rendering axis. Our *view based sampling* enables valid gradients on the x and y axis from the view.

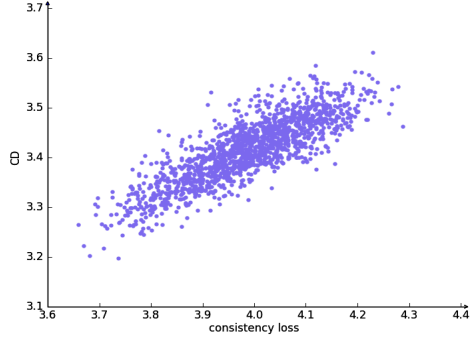Figure 8: Visualization of the multi-view reconstruction results on real world images.

Figure 9: Correlation between the consistency loss and the 3D test error.

Correlation between consistency loss and reconstruction error: In this part, we study the positive correlation between $loss_{consis}$ and the reconstruction error CD. First, we sampled $loss_{consis}$ and CD simultaneously at test stage. As shown in Figure 9, these two metrics show strong patterns of positive correlation. We further demonstrate this consistency on a highly diverse model (refer to our supplementary material for details). Table 6 shows the experimental results. Minimizing the consistency loss gives consistent decrease on the CD metric with respect to the groundtruth shape. This demonstrates the fact that with $loss_{consis}$ as the mutual constraints inside the framework, the model will infer a more accurate shape at test stage. This verifies our interpretation on the success of applying the conditional generative model to the task of multi-view shape reconstruction.

4.5. Reconstructing Real World Images

The idea of multi-view reconstruction with our conditional generative model has great generalization ability. We conducted experiments on *Stanford Online Products* dataset [27] for reconstructing real world images. Figure 8 visualizes our predictions. Our model generates surprisingly reasonable shapes by observing multi-view images in real world scenarios.

To further demonstrate the necessity of conditional modeling, Figure 10 shows visual results on unsymmetric real
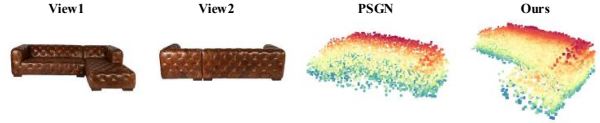

Figure 10: **Visualization on unsymmetric real-world data. Left Two:** Input images. For [6], we use both input images and then take the best prediction. For our model, we use n=2 input views. **Right Two:** While [6] tends to hallucinate the back part symmetrically, our model achieves much better results, which further demonstrates the necessity of conditional modeling.

Table 6: Study on the correlation between the consistency loss and the 3D test error CD. "heuris" denotes the heuristic search in the $\{r_i\}_{i=1}^n$ initialization. "bp" denotes the online optimization of $\{r_i\}_{i=1}^n$. "dist1" denotes GT $\rightarrow$ pred and "dist2" denotes pred $\rightarrow$ GT. Experiments on a specific model demonstrates the positive correlation between $loss_{consis}$ and CD. Both the heuristic search for initialization and the online update contribute to the performance improvement.

heuris	bp	dist1	dist2	CD	$loss_{consis}$
no	no	1.40	7.76	9.15	13.96
yes	no	1.41	3.24	4.65	5.96
yes	yes	1.40	**2.80**	**4.21**	**4.66**

data sampled from [27]. While [6] sticks to the symmetry prior and fails to generalize, our model generates a much realistic prediction.

5. Conclusion

In this paper, we have proposed a new perspective towards image-based shape generation, where we model single-view reconstruction with a partially supervised generative network conditioned on a random input. Furthermore, we present a multi-view synthesis method based on the conditional model. With the front constraint, diversity constraint and the consistency loss introduced, our method outperforms state-of-the-art approaches with interpretability. Experiments were conducted to demonstrate the effectiveness of our method. Future directions include studying the representation of the latent variables, rotation-invariant generation as well as better training strategies.

Acknowledgements

References

[1] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. Guibas. Learning representations and generative models for 3d point clouds. *arXiv preprint arXiv:1707.02392*, 2017. 4, 6

[2] A. M. N. T. H. W. B Yang, S Rosa. 3d object dense reconstruction from a single depth view with adversarial learning. In *ICCV Workshops*, 2017. 2

[3] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 2, 5, 6, 7

[4] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *ECCV*, 2016. 1, 2, 3, 6, 7

[5] B. Dai, S. Fidler, R. Urtasun, and D. Lin. Towards diverse and natural image descriptions via a conditional gan. In *ICCV*, pages 2970–2979, 2017. 2

[6] H. Fan, H. Su, and L. J. Guibas. A point set generation network for 3d object reconstruction from a single image. In *CVPR*, pages 605–613, 2017. 1, 2, 3, 5, 6, 7, 8

[7] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 4, 5

[8] T. Groueix, M. Fisher, V. G. Kim, B. Russell, and M. Aubry. AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In *CVPR*, 2018. 2

[9] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In *NIPS*, pages 5767–5777. 2017. 4

[10] J. Gwak, C. B. Choy, M. Chandraker, A. Garg, and S. Savarese. Weakly supervised 3d reconstruction with adversarial constraint. In *3DV*, 2017. 2

[11] D. Hoiem, A. A. Efros, and M. Hebert. Automatic photo pop-up. In *ACM transactions on graphics (TOG)*, volume 24, pages 577–584. ACM, 2005. 1

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017. 2

[13] A. Johnston, R. Garg, G. Carneiro, I. Reid, and A. vd Hengel. Scaling cnns for high resolution volumetric reconstruction from a single image. In *ICCV Workshops*, 2017. 2

[14] A. Kar, C. Häne, and J. Malik. Learning a multi-view stereo machine. In *NIPS*, pages 365–376. 2017. 1, 2, 3

[15] A. Kar, S. Tulsiani, J. Carreira, and J. Malik. Category-specific object reconstruction from a single image. In *CVPR*, pages 1966–1974, 2015. 1, 2

[16] S. Kim, S. Lin, S. R. JEON, D. Min, and K. Sohn. Recurrent transformer networks for semantic correspondence. In *NeurIPS*, pages 6129–6139, 2018. 2

[17] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 2

[18] A. Kurenkov, J. Ji, A. Garg, V. Mehta, J. Gwak, C. Choy, and S. Savarese. Deformnet: Free-form deformation network for 3d shape reconstruction from a single image. *arXiv preprint arXiv:1708.04672*, 2017. 2

[19] B. Lake, R. Salakhutdinov, J. Gross, and J. Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33, 2011. 2

[20] J. Li, B. M. Chen, and G. H. Lee. So-net: Self-organizing network for point cloud analysis. In *CVPR*, 2018. 4

[21] J. Li, K. Xu, S. Chaudhuri, E. Yumer, H. Zhang, and L. Guibas. Grass: Generative recursive autoencoders for shape structures. *ACM Transactions on Graphics (TOG)*, 36(4):52, 2017. 2

[22] J. J. Lim, H. Pirsiavash, and A. Torralba. Parsing ikea objects: Fine pose estimation. In *ICCV*, pages 2992–2999, 2013. 2

[23] C.-H. Lin, C. Kong, and S. Lucey. Learning efficient point cloud generation for dense 3d object reconstruction. In *AAAI*, 2018. 2, 3, 4, 6, 7

[24] N. Mellado, D. Aiger, and N. J. Mitra. Super 4pcs fast global pointcloud registration via smart indexing. In *Computer Graphics Forum*, volume 33, pages 205–215. Wiley Online Library, 2014. 2

[25] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2, 3

[26] C. Niu, J. Li, and K. Xu. Im2struct: Recovering 3d shape structure from a single rgb image. In *CVPR*, 2018. 2

[27] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, pages 4004–4012, 2016. 8

[28] E. Park, J. Yang, E. Yumer, D. Ceylan, and A. C. Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *CVPR*, pages 3500–3509, 2017. 2

[29] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NIPS*, pages 5099–5108. 2017. 4, 6

[30] R. Ranftl and V. Koltun. Deep fundamental matrix estimation. In *ECCV*, pages 284–299, 2018. 2

[31] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *IJCV*, 40(2):99–121, 2000. 3, 4

[32] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *PAMI*, 31(5):824–840, 2009. 1

[33] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 4

[34] X. Shen, H. Su, Y. Li, W. Li, S. Niu, Y. Zhao, A. Aizawa, and G. Long. A conditional variational framework for dialog generation. In *ACL*, volume 2, pages 504–509, 2017. 2

[35] D. Shin, C. C. Fowlkes, and D. Hoiem. Pixels, voxels, and views: A study of shape representations for single view 3d object shape prediction. In *CVPR*, 2018. 2

[36] X. Sun, J. Wu, X. Zhang, Z. Zhang, C. Zhang, T. Xue, J. B. Tenenbaum, and W. T. Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *CVPR*, 2018. 2

[37] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Multi-view 3d models from single images with a convolutional network. In *ECCV*, 2016. 2

[38] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *ICCV*, pages 2088–2096, 2017. 2

[39] S. Tulsiani, A. A. Efros, and J. Malik. Multi-view consistency as supervisory signal for learning shape and pose prediction. In *CVPR*, 2018. 1, 2, 5

[40] S. Tulsiani, T. Zhou, A. A. Efros, and J. Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *CVPR*, pages 2626–2634, 2017. 2

[41] S. Vicente, J. Carreira, L. Agapito, and J. Batista. Reconstructing pascal voc. In *CVPR*, pages 41–48, 2014. 1

[42] W. Wang, Q. Huang, S. You, C. Yang, and U. Neumann. Shape inpainting using 3d generative adversarial network and recurrent convolutional networks. In *ICCV*, pages 2298–2306, 2017. 2

[43] J. Wu, Y. Wang, T. Xue, X. Sun, B. Freeman, and J. Tenenbaum. Marrnet: 3d shape reconstruction via 2.5d sketches. In *NIPS*, pages 540–550. 2017. 2, 6

[44] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *NIPS*, pages 82–90. 2016. 2

[45] J. Wu, C. Zhang, X. Zhang, Z. Zhang, W. T. Freeman, and J. B. Tenenbaum. Learning shape priors for single-view 3d completion and reconstruction. In *ECCV*, 2018. 2

[46] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *WACV*, 2014. 2

[47] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *NIPS*, pages 1696–1704. 2016. 1, 2, 6

[48] L. Yu, X. Li, C.-W. Fu, D. Cohen-Or, and P.-A. Heng. Punet: Point cloud upsampling network. In *CVPR*, 2018. 2

[49] X. Zhang, Z. Zhang, C. Zhang, J. B. Tenenbaum, W. T. Freeman, and J. Wu. Learning to Reconstruct Shapes from Unseen Classes. In *NeurIPS*, 2018. 2

[50] T. Zhao, R. Zhao, and M. Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL*, volume 1, pages 654–664, 2017. 2

[51] R. Zhu, H. Kiani Galoogahi, C. Wang, and S. Lucey. Rethinking reprojection: Closing the loop for pose-aware shape reconstruction from a single image. In *ICCV*, pages 57–65, 2017. 2

[52] M. Zollhöfer, P. Stotko, A. Görlitz, C. Theobalt, M. Nießner, R. Klein, and A. Kolb. State of the art on 3d reconstruction with rgb-d cameras. In *Computer Graphics Forum*, volume 37, pages 625–652. Wiley Online Library, 2018. 2