Weakly Supervised Generative Adversarial Networks for 3D Reconstruction

JunYoung Gwak*
Stanford University
jgwak@stanford.edu

Christopher B. Choy*
Stanford University
chrischoy@ai.stanford.edu

Animesh Garg
Stanford University
garg@cs.stanford.edu

Manmohan Chandraker
NEC Laboratories America, Inc.
University of California San Diego
manu@nec-labs.com

Silvio Savarese
Stanford University
ssilvio@stanford.edu

Abstract

Volumetric 3D reconstruction has witnessed a significant progress in performance through the use of deep neural network based methods that address some of the limitations of traditional reconstruction algorithms. However, this increase in performance requires large scale annotations of 2D/3D data. This paper introduces a novel generative model for volumetric 3D reconstruction, Weakly supervised Generative Adversarial Network (WS-GAN) which reduces reliance on expensive 3D supervision. WS-GAN takes an input image, a sparse set of 2D object masks with respective camera parameters, and an unmatched 3D model as inputs during training. WS-GAN uses a learned encoding as input to a conditional 3D-model generator trained alongside a discriminator, which is constrained to the manifold of realistic 3D shapes. We bridge the representation gap between 2D masks and 3D volumes through a perspective raytrace pooling layer, that enables perspective projection and allows back-propagation. We evaluate WS-GAN on ShapeNet, ObjectNet and Stanford Online Product dataset for reconstruction with single-view and multi-view cases in both synthetic and real images. We compare our method to voxel carving and prior work with full 3D supervision. Additionally, we also demonstrate that the learned feature representation is semantically meaningful through interpolation and manipulation in input space.

1. Introduction

Recovering the three-dimensional (3D) shape of an object is a fundamental attribute of human perception. This problem has been explored by a large body of work in computer vision, within domains such as structure from motion [17, 11] or multiview stereo [12, 13, 15, 19]. While tremendous success has been achieved with conventional approaches, they often require several images to either establish accurate correspondences or ensure good coverage. This has been especially true of methods that rely on weak cues such as silhouettes [39] or aim to recover 3D volumes rather than point clouds or surfaces [25]. In contrast, human vision seems adept at 3D shape estimation from a single or a few images, which is also a useful ability for tasks such as robotic manipulation and augmented reality.

The advent of deep neural networks has allowed incorporation of semantic concepts and prior knowledge learned from large-scale datasets of examples, which has translated into approaches that achieve 3D reconstruction from a single

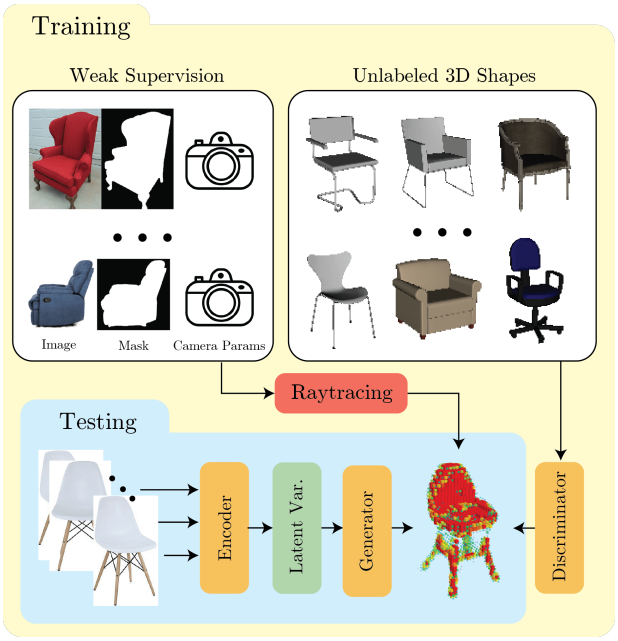
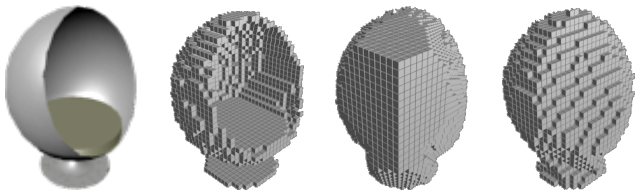


Figure 1. We use a set of unlabeled 3D shapes and images with 2D foreground segmentation mask to train the Mask Guided 3D GAN. The network generates 3D reconstruction without 3D supervision which is unavailable for many real image datasets.

*indicates equal contributions

or sparse viewpoints [5, 54, 14, 50, 51]. But conventional approaches to train convolutional neural networks (CNNs) for 3D reconstruction requires large-scale supervision. To learn the mapping from images to shapes, CAD models or point clouds are popularly used, however, ground truth alignments of models to images are challenging and expensive to acquire. Thus, existing datasets that contain image to 3D model mappings simply label the closest model as ground truth [53, 52], which leads to suboptimal training.



(a) input chair (b) voxelized input chair (c) reconst. with 1 mask (d) reconst. with 24 masks

Figure 2. Visual hull reconstruction using silhouettes is an ill-conditioned problem due to ambiguity in depth and concavity. (a) a chair with strong concavity (b) the voxelization of the corresponding chair. Visual hull reconstruction with (c) one silhouette and (d) 24 silhouettes. Note that it is not possible to reconstruct concavity using silhouettes.

This paper presents a framework for volumetric instance reconstruction using silhouettes (foreground mask) from a single or sparse set of viewpoints and corresponding camera parameters. However, visual hull reconstruction from such inputs is a severely ill-posed problem (Fig. 2). Instead, we propose constraining the space of 3D reconstruction to the to the manifold of shapes which captures the notions of plausibility, concavity, symmetry, physical stability, etc, that are difficult to capture simply using foreground masks. We implemented such constraint using projecting the reconstruction to the manifold. We explain in Sec. 3.4 that the gradient from the discriminator in the generative adversarial network (GAN) is equivalent the projection toward the manifold and the generator uses masks to synthesize shapes. To learn the perspective projection relationship between the 2D input and 3D output spaces in an end-to-end trainable framework, we propose a raytrace pooling layer in Sec. 3.3 that mimics the conventional volumetric reconstruction methods such as voxel carving [25]. Once we train the network, it only uses images at test time. An overview of our framework is in Fig. 1.

In Sec. 5, we experimentally evaluate our framework using three different datasets and report quantitative reductions in error compared with various baselines. Also, we use a synthetic dataset to generate controlled environment and vary the level of supervision, the complexity of shapes, and the number of viewpoints for ablation study. While we use GANs for reconstruction similar to [50], our work differs in requiring only relatively inexpensive 2D weak supervision as input. Further, while [54] also uses masks, we require much fewer number of them, which demonstrates better encapsulation of semantic or category-level shape information.

To summarize, the main contributions of this paper are:

- A novel framework based on deep neural networks to learn from weak supervision by exploiting the target manifold learned by an adversarial discriminator.
- Application of our framework towards 3D reconstruction from a single or few segmentation masks, avoiding the requirement of expensive 3D annotation.
- Development of a perspective voxel raytracing pooling layer to link 2D annotation with 3D representation.
- Validation on several datasets to showcase weakly supervised single view and multiview reconstruction.

2. Prior Work

In this section, we cover the related works with respect to the three aspects of our framework: Convolutional Neural Networks for 3D data, supervised 3D reconstruction, and Generative Adversarial Networks.

3D Convolutional Neural Networks. First introduced in video classification, the 3D Convolutional Neural Networks have been widely used as a tool for spatiotemporal data analysis [22, 2, 46, 30, 47]. Instead of using the third dimension for temporal convolution, [51, 27] use the third dimension for the spatial convolution and propose 3D convolutional deep networks for 3D shape classification. Recently, 3D-CNNs have been widely used for various 3D data analysis tasks such as 3D detection or classification [44, 35, 32], semantic segmentation [6, 36] and reconstruction [49, 5, 50, 14, 54]. Our work is closely related to the line of work that uses the 3D-CNN for reconstruction, as discussed in the following section.

Supervised 3D voxel reconstruction. Among many lines of work within the 3D reconstruction [17, 25, 12, 13, 3, 7, 23, 40, 49, 37], ours is related to recent works that use neural networks for 3D voxel reconstruction. Grant *et al*. [14] propose an autoencoder to learn the 3D voxelized shape embedding and regress to the embedding from 2D images using a CNN and generated 3D voxelized shape from a 2D image. Choy *et al*. [5] use a 3D-Convolutional Recurrent Neural Network to directly reconstruct a voxelized shape from multiple images of the object. [50] combine a 3D-CNN with a Generative Adversarial Network to learn the latent space of 3D shapes. Given the latent space of 3D shapes, [50] regresses the image feature from a 2D-CNN to the latent space to reconstruct a single-view image. These approaches require associated 3D shapes for training. Recently, Yan *et al*. [54] propose a way to train a neural network to reconstruct 3D shapes using a large number of foreground masks (silhouettes) and viewpoints for weak supervision. The silhouette is used to

carve out spaces analogous to voxel carving [25, 41, 28] and to generate the visual hull.

Our work is different from [54] in that it makes use of both unmatched 3D shape and inexpensive 2D weak supervision to generate realistic 3D shapes without explicit 3D supervision. This allows the network to learn reconstruction with minimal 2D supervision (as low as one view 2D mask). And the key mechanism that allows such 2D weak supervision is the projection. Unlike [54], we propose a Perspective Raytrace Pooling layer that is not limited to the grid sampling and experimentally compare with it in Sec. 5.3. In addition, we use a recurrent neural network that can handle both single and multi-view images.

Conditional Generative Adversarial Networks. First proposed by Goodfellow *et al*. [16], Generative Adversarial Networks (GAN) have been widely used for image synthesis. Recently, conditional image generation using GAN has been used for text to image synthesis [34], image inpainting [31], and image translation [20]. They design discriminators to take both input and output, letting the discriminator learn the match. Zhu *et al*. [55] use GAN to learn the manifold of artistic images and traverse the manifold given user input to manipulate an image.

Our work is similar to conditional GANs in that it generates an output conditioned on a specific input, but we do not use a discriminator that takes both input and generated output. Instead, we used the gradient from the discriminator as a projection toward the manifold of 3D shapes (Sec. 3.4).

3. Weakly Supervised GAN

Recent supervised single view reconstruction methods [14, 5, 49, 50] require associated 3D shapes. However, such 3D annotations are hard to acquire for real image datasets such as [8, 43]. Instead, we propose a framework, termed as Weakly supervised Generative Adversarial Network (WS-GAN), that relies on inexpensive 2D silhouette and approximate viewpoint for weak supervision. WS-GAN makes use of unlabeled 3D shapes to constrain the ill-posed single/sparse-view reconstruction problem. Specifically, we use Generative Adversarial Network to make use of unlabeled 3D shapes which guide the ill-posed reconstruction. To make use of 2D weak supervision, we propose a Perspective Raytrace Pooling layer. Following sections consist of the overview of the framework (Sec. 3.1), encoding images and reconstruction (Sec. 3.2), Raytracing (Sec. 3.3), and optimization (Sec. 3.5).

3.1. WS-GAN Overview

WS-GAN consists of four components: an encoder $E(\cdot)$ that maps images I to latent variable z; a generator $G(\cdot)$ that generates the voxelized reconstruction x from the latent variable; a PRP-Layer that takes both viewpoint c_i and reconstruction x and generate rendering; and a discriminator that takes either voxelized unlabeled 3D shapes $\hat{x}$ or voxelized reconstruction x and returns a scalar value. During the training, we use an annotated image dataset $\mathcal{D} = \{(\mathbf{I}_i, m_i, c_i)\}_i$ that consists of an image I, a corresponding foreground mask (silhouette) m, and the approximate viewpoint c. At test time, the model will only have access to input images I, and the required output would be 3D reconstruction x. Note that we use a Recurrent Reconstruction Neural Network [5] for the base network, which allows WS-GAN to take multi-view images as well if available. Complete diagram is illustrated in the Fig. 3.

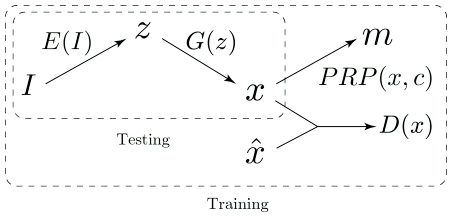


Figure 3. WS-GAN takes as input I, m, c, I is the image, m is the foreground segmentation mask and c is camera extrinsics. An encoded representation z is first learned which guides the generator G to output a 3D reconstruction x. Given the set $\mathcal{D}_S$ of unlabeled 3D shapes $\hat{x}$, the discriminator $D(\cdot)$ outputs the probability $p(x)$ of x being realistic.

3.2. Encoder and Generator

Given single/multi-view images of an instance $\mathbf{I} = \{I_1, I_2, \ldots, I_M\}$, the encoder E encodes images into a latent variable z. We used a recurrent neural network with 3D Convolutional GRU for the encoder. The GRU provides attention and to map an image feature into a latent variable that has 3D spatial structure. Next, the generator G decodes the latent variable into 3D reconstruction x. The 3D reconstruction is in the form of probabilistic voxel occupancy map where each voxel has occupancy probability. The reconstruction $x \in \mathbb{R}^{N_v \times N_v \times N_v}$, where N_v is the reconstruction resolution. The entire process can be summarized as $G(E(\mathbf{I})) \sim q(x|I)$ and $q(x|I)$ is the conditional distribution. Following [5], we used a 2D Convolutional Neural Network for the encoder and a 3D Convolutional Neural Network for the generator. Both networks have residual connections to speed up the convergence and improve performance [18].

3.3. Perspective Raytrace Pooling

The 2D weak supervisions reside in the image domain whereas the reconstruction is in 3D space. To bridge different domains, we propose a Perspective Raytrace Pooling layer (PRP-Layer). It takes a 3D volumetric reconstruction x and camera viewpoint c and generates the rendering of the reconstruction x. c is composed of camera center, C and camera perspective R. Let a ray emanating from camera

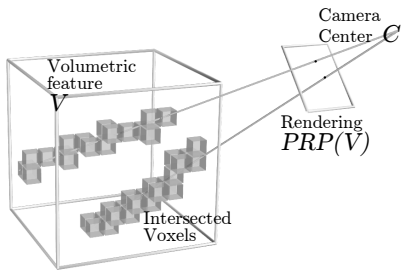


Figure 4. Visualization of raytrace pooling. For each pixel of 2D rendering, we calculate the direction of the ray from camera center. Then, we apply pooling function to all hit voxels in 3D grid.

center C be $\mathbf{L}_i$ and the intersection of the ray with the image plane be p_i. Then, ray can be parametrized by $t \in \mathbb{R}_+$

$$\mathbf{L}(t) = \mathbf{C} + t\frac{\mathbf{R}^{-1}p - \mathbf{C}}{\|\mathbf{R}^{-1}p - \mathbf{C}\|} \tag{1}$$

We aggregate all the voxels v_j that intersect with the ray $\mathbf{L}_i$ using an octree voxelwalking [1] with efficient ray-box intersection algorithm [48], and compute a single feature for each ray f_i by pooling over the features in the voxels. We visualized the result of the raytracing and aggregated voxels in Fig. 4. While multiple types of pooling operations are admissible, we use max pooling in this work. Max pooling along the ray $\mathbf{L}_i$ in an occupancy grid x results in a foreground mask $\tilde{m}$.Finally, we can measure the difference between the predicted foreground mask $\tilde{m} = PRP(x, c_j)$ and the ground truth foreground mask m and define a loss $\mathcal{L}_{\text{ray}}$.

$$\mathcal{L}_{\text{ray}}(x, \mathbf{c}, m) = \mathop{\mathbb{E}}_{x\sim q}\left[\frac{1}{M}\sum_{j}^{M}\mathcal{L}_{\text{s}}(PRP(x, c_j), m_j)\right],$$

where M is the number of silhouettes from different viewpoints and c_j is the j-th the camera viewpoint, and $\mathcal{L}_{\text{s}}$ is the cross-entropy loss. Instead of using raytracing for rendering, a concurrent work in [54] has independently proposed a projection layer based on the Spatial Transformer Network [21]. Since the sampling can cause aliasing if the sampling rate is lower than the Nyquist rate [29], the sampling grid from [54] has to be dense and compact. On the other hand, PRP-Layer mimics the rendering process and does not suffer from aliasing. To see the effect of aliasing of sampling based projection, we compare the performance of [54] and PRP-Layer in Sec. 5.3.

3.4. Discriminator

The discriminator $D(\cdot) \in [0, 1]$ takes in the 3D voxel occupancy map and returns score that measures the likelihood of the input being drawn from real distribution $p(x)$. From the proposition 1 in [16], the optimal discriminator given G is $D^*(x) = \frac{p(x)}{p(x)+q(x)}$. Intuitively, the gradient of $D(x)$ with respect to x, $\frac{\partial D(x)}{\partial x}$, gives the local direction toward higher $p(x)$. This gives a direction toward the manifold of realistic 3D shapes and forces the generator output to constrain the reconstruction to realistic 3D shape space. In other words, if the 2D weak supervision only provides visual hull, the discriminator can force generator to recover concavity, symmetry, unseen parts, etc by making use of the unlabeled 3D shapes. We illustrated the distribution of 3D shape space and the gradient from the discriminator in Fig. 5.

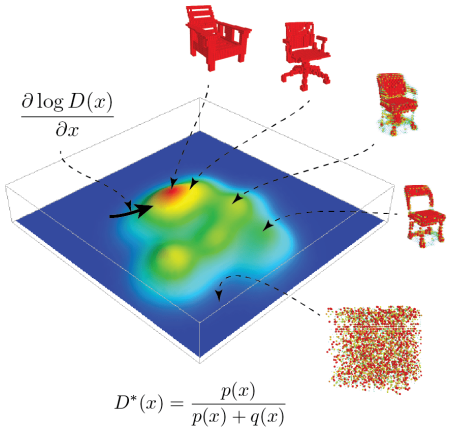


Figure 5. Illustration of the discriminator score. $D(x)$ learns the manifold of 3D realistic shapes and the gradient $\partial \log D(x)/\partial x$ projects the generated shape $G(x|\mathbf{I})$ toward the space of 3D shapes.

3.5. WS-GAN Optimization

Finally, we formulate the problem as a minimax GAN optimization problem

$$\min_{\theta_G,\theta_E}\max_{\theta_D}\mathcal{L}(x) + \lambda \cdot \mathcal{L}_{ray}(x, c, m) \tag{2}$$

$$\mathcal{L}(x) = \mathop{\mathbb{E}}_{\hat{x}\sim p}\log D(\hat{x}) + \mathop{\mathbb{E}}_{x\sim q}\log(1 - D(x)) \tag{3}$$

where $\mathcal{L}_{ray}$ is the weak supervision loss from silhouettes m and approximate viewpoint c and $p(x)$ is the probability distribution of the unlabeled 3D shapes and q denotes the probability distribution of reconstruction $q(x|\mathbf{I})$. The final algorithm is in Algo. 1.

While convergence properties of such an optimization problem are non-trivial to prove and an active area of research, our empirical results consistently indicate it behaves reasonably well in practice.

4. Implementation

Network We implemented WS-GAN with a symbolic neural network library [45]. For the generator, we use the Deep-Residual-GRU network proposed by Choy *et al.* [5], which uses a 2D convolutions with and 3D decoder with residual connections across layers and GRUs for RNN implementation. The schematics of WS-GAN is in Fig. 6. We used RGB images of size 127^2 as an input and the voxel reconstruction of size 32^3 ($N_v = 32$) as an output. For the PRP-Layer, we use the silhouette of size 127^2. Finally, for the discriminator, we stacked 3D convolutions and 3D poolings until the size of activation becomes $2 \times 2 \times 2$ and flatten to feed

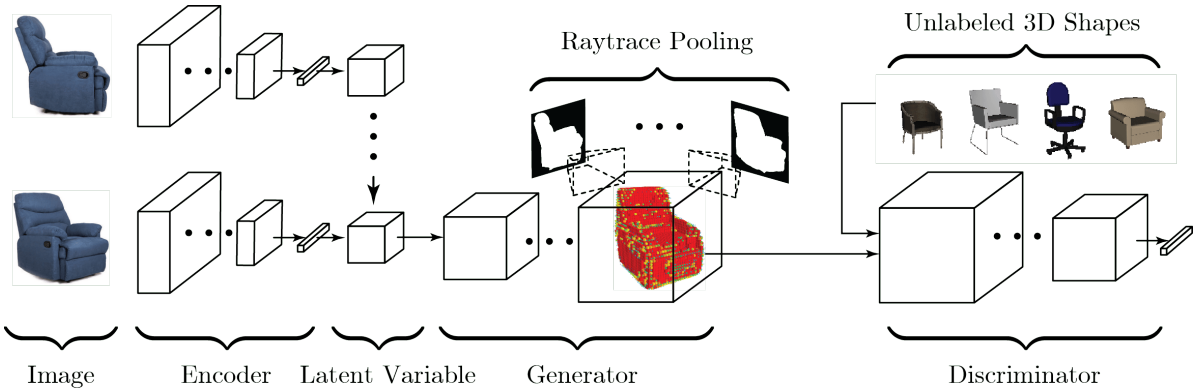


Figure 6. Visualization of WS-GAN network structure. Our network encodes a set of images into a latent variable. Then, the latent variable is decoded into a voxel representation of 3D shape. Perspective Raytrace Pooling layer renders this 3D shape into 2D occupancy map, allowing us to give mask supervision. Additionally, discriminator takes the generated voxel as an input, filliing the missing information of the 3D shape distribution learned from unlabeled 3D models.

Algorithm 1 Weakly supervised Generative Adversarial Network: Training

Require: Datasets: $\mathcal{D} = \{(\mathbf{I}_i, m_i, c_i)\}, \mathcal{D}_S = \{\hat{x}\}$

while not converged **do**
  for all images $(\mathbf{I}_i, m_i, c_i) \in \mathcal{D}$ **do**
    $z \leftarrow E(\mathbf{I}_i)$ // Encode images
    $x \leftarrow G(z)$ // Generate 3D model
    for all camera $c_{i,j}$, s.t. $j \in \{1, \cdots, M\}$ **do**
      $\tilde{m}_{i,j} \leftarrow PRP(x, c_{i,j})$ // Perspective Proj.
    end for
    $\mathcal{L}_{ray} \leftarrow \frac{1}{M} \sum_{j=1}^{M} \mathcal{L}_s(\tilde{m}_{i,j}, m_{i,j})$
    $\mathcal{L}_G \leftarrow \mathcal{L}(D(x), G) + \lambda \mathcal{L}_{ray}$ // Generator loss
    $\theta_G \leftarrow \theta_G - \alpha \partial \mathcal{L}_G / \partial \theta_G$ // Grad. descent for G
    $\theta_E \leftarrow \theta_E - \alpha \partial \mathcal{L}_G / \partial \theta_E$ // Grad. descent for E
    $\mathcal{L}_D \leftarrow \mathcal{L}(D(\hat{x}), G)$ // Discriminator loss
    $\theta_D \leftarrow \theta_D + \beta \partial \mathcal{L}_D / \partial \theta_D$ // Grad. ascent for D
  end for
end while

the activations to a fully connected layer followed by a softmax layer. The details of the network architecture are in the supplementary material.

Optimization Training GAN is notoriously difficult [33, 38, 42] and training the Weakly Supervised GAN for 3D Reconstruction was not an exception. Inherently, GAN training involves computing $\log q/p$ which can cause divergence if support of p does not overlap with the support of q. To prevent such case, we followed the instance noise technique by Sønderby *et al.* [42] which smooths the probability space to make the support of p infinite. In addition, we used the update rule by Wu *et al.* [50] and train the discriminator only if its prediction error becomes larger than 20%. This technique makes the discriminator imperfect and prevents saturation of D. Finally, we use different learning rate for the discriminator and the generator: 10^{-2} for θ_G and 10^{-4} for θ_D and reduce the learning rate by the factor of 10 after 10,000 and 30,000 iterations. We train the network over 40,000 iterations using ADAM [24] with batch size 8. We used loss balancing hyper-parameter $\lambda = 0.01$ for all experiments.

5. Experiments

To validate our approach, we designed various experiments and used standard datasets. First, we define the baseline methods including recent works (Sec. 5.1) and evaluation metrics (Sec. 5.2). To compare our approach with baseline methods in a controlled environment, we used a 3D shape dataset and rendering images. We present quantitative ablation study results on Sec. 5.3. Next, we test our framework on a real image single-view and a multi-view dataset in Sec. 5.4 and Sec. 5.5 respectively. To examine the representation of the latent variable z, we follow [33, 50] and analyze the semantic representation of the latent variable (Sec 5.6) similar to [33, 50]. Note that, we can manipulate the output (shapes) using different modality (image) and allow editing in a different domain.

5.1. Baselines

For accurate ablation study, we propose various baselines to examine each component in isolation. First, we categorize all the baseline methods into three categories based on the level of supervision: *2D Weak Supervision (2D)*, *2D Weak Supervision + unlabeled 3D Supervision (2D + U3D)*, and *Full 3D Supervision (F3D)*. *2D* has access to 2D silhouettes and viewpoints as supervision; and *2D + U3D* uses silhouettes, viewpoints, and unlabeled 3D shapes for supervision. Finally, *F3D* is supervised with the ground truth 3D reconstruction associated with the images. Given *F3D* supervision, silhouettes do not add any information, thus the performance of a system with full supervision provides an approximate performance upper bound.

Specifically, in the *2D* case, we use Perspective Raytrace

IOU / AP

Level of supervision	Methods	Transportation		Furniture				Mean
		car	airplane	sofa	chair	table	bench	
1 view 2D	VC [25]	0.2605 / 0.2402	0.1092 / 0.0806	0.2627 / 0.2451	0.2035 / 0.1852	0.1735 / 0.1546	0.1303 / 0.1064	0.1986 / 0.1781
	PTN [54]	0.4437 / 0.7725	0.3352 / 0.5568	0.3309 / 0.4947	0.2241 / 0.3178	0.1977 / 0.2800	0.2145 / 0.2884	0.2931 / 0.4620
	PRP	0.3791 / 0.7250	0.2508 / 0.4997	0.3427 / 0.5093	0.1930 / 0.3361	0.1821 / 0.2664	0.2188 / 0.3003	0.2577 / 0.4452
1 view 2D + U3D	PRP+NN	0.5451 / 0.5582	0.2057 / 0.1560	0.2767 / 0.2285	0.1556 / 0.1056	0.1285 / 0.0872	0.1758 / 0.1183	0.2597 / 0.2267
	WSGAN	**0.5622 / 0.8244**	**0.3727 / 0.5911**	**0.3791 / 0.5597**	**0.3503 / 0.4828**	**0.3532 / 0.4582**	**0.2953 / 0.3912**	**0.4036 / 0.5729**
5 views 2D	VC [25]	0.5784 / 0.5430	0.3452 / 0.2936	0.5257 / 0.4941	0.4048 / 0.3509	0.3549 / 0.3011	0.3387 / 0.2788	0.4336 / 0.3857
	PTN [54]	0.6593 / 0.8504	0.4422 / 0.6721	0.5188 / 0.7180	0.3736 / 0.5081	0.3556 / 0.5367	0.3374 / 0.4725	0.4572 / 0.6409
	PRP	0.6521 / **0.8713**	0.4344 / 0.6694	0.5242 / 0.7023	0.3717 / 0.5048	0.3197 / 0.4464	0.321 / 0.4377	0.4442 / 0.6123
5 views 2D + U3D	PRP+NN	**0.6744** / 0.6508	**0.4671** / 0.4187	0.5467 / 0.5079	0.3449 / 0.2829	0.3081 / 0.2501	0.3116 / 0.2477	0.4465 / 0.3985
	WS-GAN	0.6142 / 0.8674	0.4523 / **0.6877**	0.5458 / **0.7473**	**0.4365 / 0.6212**	**0.4204 / 0.5741**	**0.4009 / 0.5770**	**0.4849 / 0.6851**
F3D	R2N2 [5]	0.8338 / 0.9631	0.5425 / 0.7747	0.6784 / 0.8582	0.5174 / 0.7266	0.5589 / 0.7754	0.4950 / 0.6982	0.6210 / 0.8123

Table 1. Per-category 3D reconstruction Intersection-over-Union(IOU) / Average Precision(AP). Please see Sec. 5.1 for details of baseline methods and the level of supervision. WS-GAN outperforms other baselines by larger margin in classes with more complicated shapes as shown in Fig. 7.

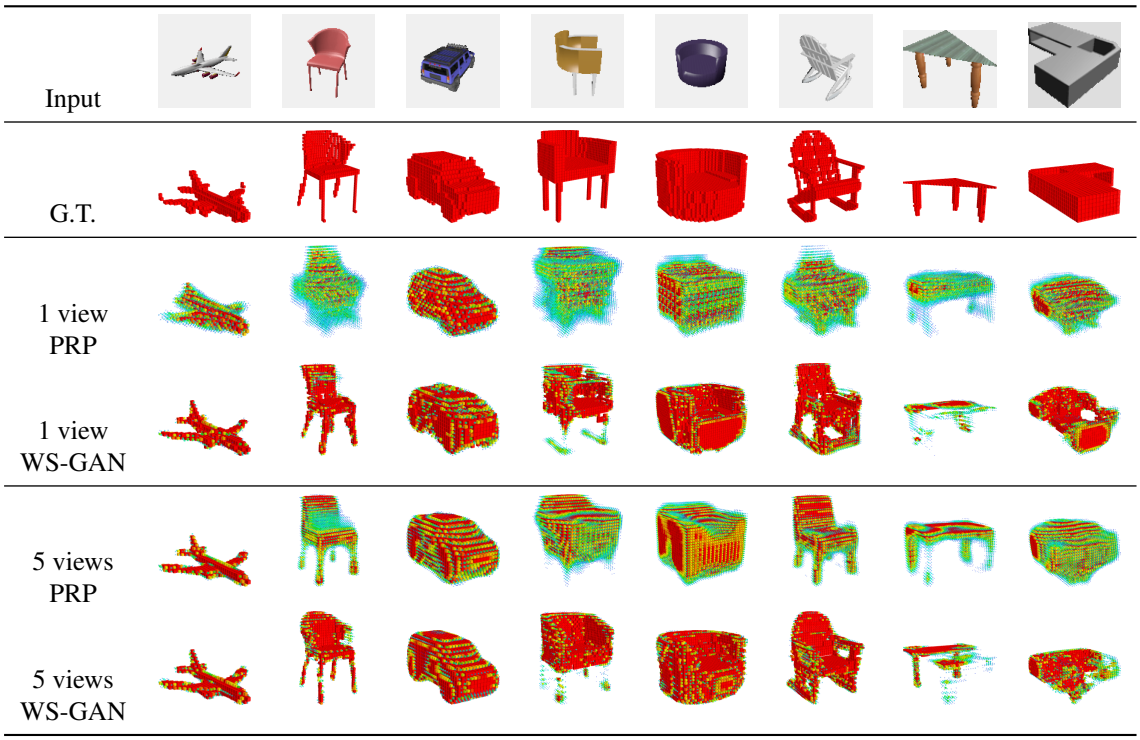


Figure 7. Qualitative results of single- or multi-view synthetic image reconstructions on ShapeNet dataset. Compared to voxel carving and PRP which only uses 2D weak supervision, WS-GAN reconstructs complex shapes better. Please refer to Sec. 5.2 for details of our visualization method.

Pooling (PRP) as proposed in Sec .3.3 and compare it with Perspective Transformer (PTN) by Yan *et al*. [54]. Next, in the *2D + U3D* case, we use PRP + Nearest Neighbor (PRP+NN) and WS-GAN. PRP + NN uses unlabeled 3D shapes, by retrieving the 3D shape that is closest to the prediction. Finally, in the *F3D* case, we use R2N2 [5]. We did not include [50, 14] in this experiment since they are restricted to single-view reconstruction and use full 3D supervision which would only provide an additional upper bound. For all neural network based baselines, we used the same base network architecture (encoder and generator) to ascribe performance gain only to the supervision mode. Aside from learning-based methods, we also provide a lower-bound on performance using voxel carving (VC) [25]. We note that voxel carving requires silhouette and camera viewpoint during testing. Kindly refer to the supplementary material for details of baseline methods, implementation, and training.

5.2. Metrics and Visualization

The network generates a voxelized reconstruction, and for each voxel, we have occupancy probability (confidence). We use Average Precision (AP) to evaluate the quality and

the confidence of the reconstruction. We also binarize the probability and report Intersection-over-Union (IOU) with threshold 0.4, following [5]. This metric gives more accurate evaluation of deterministic methods like voxel carving. For visualization, we use red to indicate voxels with occupancy probability above 0.6 and gradually make it smaller and green until occupancy probability reaches 0.1. When the probability is below 0.1, we did not visualize the voxel.

5.3. Ablation Study on ShapeNet [4]

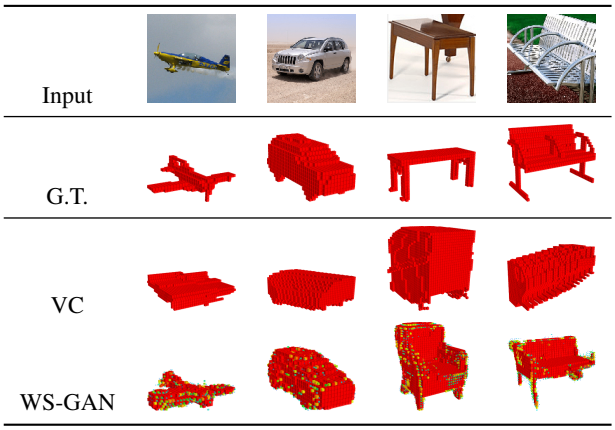


Figure 8. Real image single-view reconstructions on ObjectNet3D. Compared to PRP which only uses 2D weak supervision, WS-GAN reconstructs complex shapes better. Please refer to Sec. 5.2 for details of our visualization method.

Figure 9. Qualitative results of multi-view real image reconstructions on Stanford Online Product dataset [43]. Our network successfully reconstructed real images coordinating multi-view information.

In this section, we perform ablation study and compare WS-GAN with the baseline methods on the ShapeNet [4] dataset. The synthetic dataset allows us to control external factors such as the number of viewpoints, quality of mask and is ideal for ablation study. Specifically, we use the renderings from [5] since it contains a large number of images from various viewpoints and the camera model has more degree of freedom. In order to train the network on multiple categories while maintaining a semantically meaningful

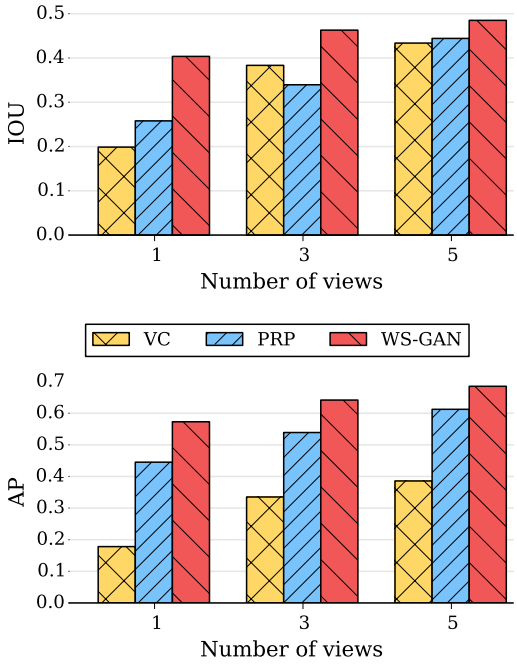


Figure 10. Intersection-over-union (IOU) and Average Precision (AP) over the number of masks used for weak supervision. The performance gap between WS-GAN and the other baselines gets larger as the number of views of masks decreases (i.e. supervision strength gets weaker).

manifold across different classes, we divide the categories into furniture (sofa, chair, bench, table) and vehicles (car, airplane) classes and trained networks separately. We use the alpha channel of the renderings image to generate 2D mask supervisions (finite depth to indicate foreground silhouette). For the unlabeled 3D shapes, we simply voxelized the 3D shapes. To simulate realistic scenario, we divide the dataset into three **disjoint** sets: shapes for 2D weak supervision, shapes for unlabeled 3D shapes, and the test set. Next, we study the impact of the level of supervision, the number of viewpoints, and the object category on the performance.

First, we found that more supervision leads to better reconstruction and WS-GAN make use of the unlabeled 3D shapes effectively (Vertical axis of Tab. 1). Compare with the simple nearest neighbor, which also make use of the unlabeled 3D data, WS-GAN outperforms the simple baseline by a large margin. This hints that WS-GAN smoothly interpolates the manifold of 3D shapes and project the 3D reconstruction toward the manifold as we conjectured in Sec. 3.4. Second, WS-GAN learns to generate better reconstruction even from a small number of 2D weak supervision. In Tab. 1 and in Fig. 10, we vary the number of 2D silhouettes that we used to train the networks and observe that the performance improvement that we get from exploiting the unlabeled 3D shapes gets larger as we use a fewer number

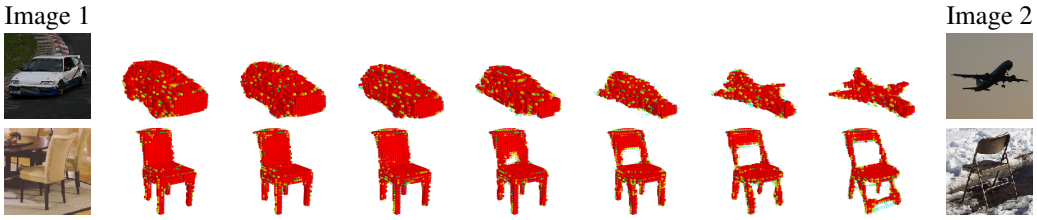


Figure 11. Linear interpolation of latent variable z. We observed the smooth transition of objects inter-and intra-class. Interestingly, semantic properties of the object, such as the length of the airplane wings and the size of the hole in the back of the chair smoothly transitioned. This result hints that our network generalized such semantic properties in the latent variable z.

	sofa	chair	table	bench	mean
PTN_16 [54]	0.4753	0.2888	0.2476	0.2576	0.2979
PTN_32 [54]	0.4947	0.3178	0.2800	0.2884	0.3283
PTN_64 [54]	0.5082	0.3377	0.3114	0.3104	0.3509
PTN_128 [54]	0.5217	0.3424	0.3104	0.3146	0.3545
PRP	0.5093	0.3361	0.2664	0.3003	0.3308

Table 2. AP of 2D weak supervision methods on single-view furniture reconstruction. In order to analyze the effect of aliasing of PTN [54], we varied its disparity sampling density (sampling density N, for all PTN_N) and compare with PRP.

of 2D supervision. Third, we observed that WS-GAN outperforms other baselines by a larger margin on classes with more complicated shapes such as chair, bench, and table which have concavity that is difficult to recover only using 2D silhouettes. For categories with simpler shapes such as car, the marginal benefit of using the adversarial network is smaller. Similarly, 3D nearest neighbor retrieval improves reconstruction quality only on few categories of a simple shape such as car while it also harms the reconstruction on complex shapes such as chair or table. This is expected since their 3D shapes are close to convex shapes and 2D supervision is enough to recover 3D shapes.

We visualized of the reconstructions in Fig. 7. We can observe that our network is capable of carving concavity, which is difficult to learn solely from mask supervision. Also, compared to the network trained only using mask supervision, WS-GAN prefers to binarize the occupancy probability, which seems to be an artifact of generator fooling discriminator.

Raytracing Comparison In this section, we compare a raytracing based projection (PRP-Layer) and a sampling based projection (PTN [54]) experimentally on ShapeNet single view furniture category. We only vary the projection method and sampling rate along depth but keep the same base network architecture. As shown in Table. 2, the reconstruction performance improves as the sampling rate increases as expected in Sec. 3.3. We suspect that the trilinear interpolation in PTN played a significant role after it reaches resolution 64 and that implementing a similar scheme using ray length in PRP-Layer could potentially improve the result.

5.4. Single-view reconst. on ObjectNet3D [52]

In this experiment, we train our network for single real-image reconstruction using ObjectNet3D [52] dataset. The dataset contains 3D annotations in the form of the closest 3D shape from ShapeNet and viewpoint alignment. Thus, we generate 2D silhouettes using 3D shapes. We split the dataset using the shape index to generate disjoint sets like the previous experiments. Since the dataset consists of at most 1,000 instances per category, we freeze the generator, and discriminator and fine-tune only the 2D encoder $E(u)$. The quantitative evaluation is available in the supplementary material and qualitative results are available in Fig. 8.

5.5. Multi-view Reconst. on OnlineProduct [43]

Stanford Online Product [43] is a large-scale multi-view dataset consisting of images of products from e-commerce websites. In this experiment, we test WS-GAN on multi-view real images using the network trained on the ShapeNet [4] dataset with random background images from PASCAL [10] to make the network robust to the background noise. We visualize the results in Fig. 9. The result shows that our network can integrate information across multiple views of real images and reconstruct a reasonable 3D shape.

5.6. Representation analysis

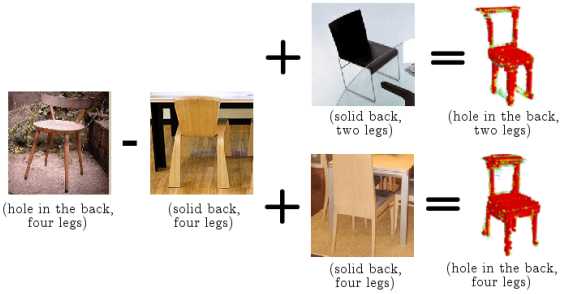


Figure 12. Arithmetic on latent variable z of different images. By subtracting latent variables of similar chairs with different properties, we extracted the feature which represents such property. We applied the feature to two other chairs to demonstrate that this is a generic and replicable representation.

In this experiment, we explore the semantic expressiveness of the latent variable z using interpolation and vector arithmetic similar to [9, 33, 50, 14]. However, unlike the

above approaches, we use different modalities for the input and output which are images and 3D shapes respectively. Therefore, we can make high-level manipulation of the latent variable $z = E(\mathbf{I})$ from 2D images and modify the output 3D shape.

First, we linearly interpolate the latent variables from two images inter-and intra-class (Fig. 11). We observed that the transition is smooth across various semantic properties of the 3D shapes such as length of the wing and the size of the hole on the back of the chair. Second, we extract a latent vector that contains semantic property (such as making a hole in a chair) and apply it on a different image to modify the reconstruction (Fig. 12).

6. Conclusion

We have proposed a novel framework that allows generative adversarial networks to reconstruct 3D shapes from sparse 2D silhouettes and unlabeled 3D shapes. To make this framework possible, we proposed perspective raytrace pooling layer which allows end-to-end training across 2D and 3D domain of information. Our network generated a high-quality reconstruction comparable to the other works with full 3D supervision. We also analyzed the shape manifold the network learned through latent variable manipulation.

References

[1] J. Arvo. Linear-time voxel walking for octrees. *Ray Tracing News*, 1(2), 1988. 4

[2] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential deep learning for human action recognition. In *Proceedings of the Second International Conference on Human Behavior Unterstanding*. Springer-Verlag, 2011. 2

[3] Y. Bao, M. Chandraker, Y. Lin, and S. Savarese. Dense object reconstruction using semantic priors. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 2

[4] 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], 2015. 7, 8, 12

[5] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 2, 3, 4, 6, 7, 11

[6] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *arXiv preprint arXiv:1702.04405*, 2017. 2

[7] A. Dame, V. A. Prisacariu, C. Y. Ren, and I. Reid. Dense reconstruction using 3d object shape priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1288–1295, 2013. 2

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 3

[9] A. Dosovitskiy, J. Tobias Springenberg, and T. Brox. Learning to generate chairs with convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 8

[10] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, Jan. 2015. 8, 11, 12

[11] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha. Visual simultaneous localization and mapping: a survey. *Artificial Intelligence Review*, 43, 2015. 1

[12] Y. Furukawa, B. Curless, S. Seitz, and R. Szeliski. Towards internet-scale multi-view stereo. In *CVPR*, pages 1434–1441, 2010. 1, 2

[13] Y. Furukawa and J. Ponce. Accurate, dense and robust multi-view stereopsis. *PAMI*, 32(8):1362–1376, 2010. 1, 2

[14] R. Girdhar, D. Fouhey, M. Rodriguez, and A. Gupta. Learning a predictable and generative vector representation for objects. In *ECCV*, 2016. 2, 3, 6, 8, 11

[15] M. Goesele, J. Ackermann, S. Fuhrmann, R. Klowsky, F. Langguth, P. Müandcke, and M. Ritz. Scene reconstruction from community photo collections. *IEEE Computer*, 43:48–53, 2010. 1

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014. 3, 4

[17] K. Häming and G. Peters. The structure-from-motion reconstruction pipeline–a survey with focus on short image sequences. *Kybernetika*, 46(5):926–937, 2010. 1, 2

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[19] C. Hernández and G. Vogiatzis. Shape from photographs: A multi-view stereo pipeline. In *Computer Vision*, volume 285 of *Studies in Computational Intelligence*, pages 281–311. Springer, 2010. 1

[20] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arxiv*, 2016. 3

[21] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu. Spatial transformer networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2017–2025. Curran Associates, Inc., 2015. 4

[22] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. In *ICML*, 2010. 2

[23] A. Kar, S. Tulsiani, J. Carreira, and J. Malik. Category-specific object reconstruction from a single image. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1966–1974. IEEE, 2015. 2

[24] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[25] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. *International Journal of Computer Vision*, 38(3):199–218, 2000. 1, 2, 3, 6, 11

[26] J. J. Lim, H. Pirsiavash, and A. Torralba. Parsing IKEA Objects: Fine Pose Estimation. *ICCV*, 2013. 11

[27] D. Maturana and S. Scherer. VoxNet: A 3D Convolutional Neural Network for Real-Time Object Recognition. In *IROS*, 2015. 2

[28] W. Matusik, C. Buehler, R. Raskar, S. J. Gortler, and L. McMillan. Image-based visual hulls. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 369–374. ACM Press/Addison-Wesley Publishing Co., 2000. 3

[29] A. V. Oppenheim and R. W. Schafer. *Discrete-Time Signal Processing*. Prentice Hall Press, Upper Saddle River, NJ, USA, 3rd edition, 2009. 4

[30] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui. Jointly modeling embedding and translation to bridge video and language. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[31] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 3

[32] C. R. Qi, H. Su, M. Niessner, A. Dai, M. Yan, and L. J. Guibas. Volumetric and multi-view cnns for object classification on 3d data. *arXiv preprint arXiv:1604.03265*, 2016. 2

[33] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 5, 8

[34] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text-to-image synthesis. In *Proceedings of The 33rd International Conference on Machine Learning*, 2016. 3

[35] Z. Ren and E. B. Sudderth. Three-dimensional object detection and layout prediction using clouds of oriented gradients. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[36] G. Riegler, A. Osman Ulusoy, and A. Geiger. OctNet: Learning Deep 3D Representations at High Resolutions. *ArXiv e-prints*, Nov. 2016. 2

[37] J. Rock, T. Gupta, J. Thorsen, J. Gwak, D. Shin, and D. Hoiem. Completing 3d object shape from one depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2484–2493, 2015. 2

[38] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved Techniques for Training GANs. *ArXiv e-prints*, 2016. 5

[39] S. Savarese, M. Andreetto, H. Rushmeier, F. Bernardini, and P. Perona. 3d reconstruction by shadow carving: Theory and practical evaluation. *International Journal of Computer Vision*, 71(3):305–336, 2007. 1

[40] N. Savinov, C. Häne, M. Pollefeys, et al. Discrete optimization of ray potentials for semantic 3d reconstruction. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5511–5518. IEEE, 2015. 2

[41] S. M. Seitz and C. R. Dyer. Photorealistic scene reconstruction by voxel coloring. *International Journal of Computer Vision*, 35(2):151–173, 1999. 3

[42] C. K. Sønderby, J. Caballero, L. Theis, W. Shi, and F. Huszár. Amortised map inference for image super-resolution. *arXiv preprint arXiv:1610.04490*, 2016. 5

[43] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 7, 8, 12, 15

[44] S. Song and J. Xiao. Deep Sliding Shapes for amodal 3D object detection in RGB-D images. In *CVPR*, 2016. 2

[45] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016. 4

[46] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15. IEEE Computer Society, 2015. 2

[47] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Deep end2end voxel2voxel prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016. 2

[48] A. Williams, S. Barrus, R. K. Morley, and P. Shirley. An efficient and robust ray-box intersection algorithm. In *ACM SIGGRAPH 2005 Courses*, page 9. ACM, 2005. 4

[49] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. Single Image 3D Interpreter Network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 2, 3

[50] J. Wu, C. Zhang, T. Xue, W. T. Freeman, and J. B. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Neural Information Processing Systems (NIPS)*, 2016. 2, 3, 5, 6, 8, 11

[51] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[52] Y. Xiang, W. Kim, W. Chen, J. Ji, C. Choy, H. Su, R. Mottaghi, L. Guibas, and S. Savarese. Objectnet3d: A large scale database for 3d object recognition. In *European Conference on Computer Vision*, pages 160–176. Springer, 2016. 2, 8, 12

[53] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision*, pages 75–82. IEEE, 2014. 2

[54] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee. Learning volumetric 3d object reconstruction from single-view with projective transformations. In *Advances in Neural Information Processing Systems*, 2016. 2, 3, 4, 6, 8, 11

[55] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros. Generative visual manipulation on the natural image manifold. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016. 3

A.1. Detailed network architecture

The network is composed of three parts - an encoder, a generator, and a discriminator. Please refer to Fig. A1 for detailed visualization of the network architectures. All source code and models used in this work will be publicly released upon publication.

First, the encoder takes RGB image(s) $\mathbf{I}$ with size 127^2 as an input. Each of the multi-view images is encoded into a feature vector of size 1024 through a sequence of convolutions and pooling with residual connections. The encoded feature vectors are reduced into a latent variable z of size $4 \times 4 \times 4 \times 128$ through 3D convolutional LSTM [5]. The first three dimensions indicate the three spatial dimensions and the last dimension indicates the feature size. The 3D convolutional LSTM works as an attention mechanism that writes features from images to corresponding voxels in 3D space. Thus, the 3D-LSTM explicitly resolves the viewpoints and self-occlusion. The encoder network is visualized in Fig. A1 (a).

Second, as shown in Fig. A1 (b), the generator repeats 3D convolution and unpooling until it reaches the resolution $32 \times 32 \times 32$ with residual connections like the encoder. Then, we apply one convolution followed by a softmax function to generate 3D voxel occupancy map x. Given the reconstruction, we compute the projection loss using the Raytrace Pooling Layer.

Lastly, the discriminator takes either the reconstruction or the unlabeled shapes and generates a scalar value. The discriminator consists of a sequence of 3D convolutions and 3D max pooling until the activation is reduced to $2 \times 2 \times 2$ grid. The activation is then vectorized and fed into a fully connected layer followed by a softmax layer. Again, the network's detailed structure can be found in Fig. A1 (c).

A.2. Baseline Methods

In this section, we cover further implementation details of the baseline methods used in the main paper.

Voxel Carving (VC): Given silhouettes and camera parameters, voxel carving [25] removes voxels that lie outside of the silhouettes when projected to the image planes. Please note that voxel carving always requires camera parameters and masks, in contrast to all other learning-based methods which only require an image as an input.

Perspective Raytrace Pooling (PRP): We train an encoder-generator network only with mask supervisions ($\mathcal{L}_{ray}$). The network has the same architecture as the WS-GAN as shown in Sec. A.1 but does not have a discriminator that provides gradients toward 3D shape manifold. Please note that the mask supervision requires Perspective Raytrace Pooling that we proposed.

Perspective Transformer (PT): [54] proposed a perspective projection layer (Perspective Transformer) that is similar to the PRP Layer. To compare it with the PRP Layer, we propose another baseline, an encoder-generator network only with mask supervisions, but with the Perspective Transformer (PT). Since the base network architecture affects the performance drastically, we use the same network for all learning based methods including this one. While the PRP uses an accurate raytracing, the PT uses sampling points from a 3D grid over a fixed range of depth from camera center on the voxel space. Therefore, the PT requires hyperparameters for the range and the density of the samples. We determined the range by experimentally measuring the minimal and maximal possible depth of the voxel space over the training data and used sampling density 16 by default as suggested by [54]. Additionally, we vary the density of the sample to measure the effect of the sampling at Sec. 5.3.

Perspective Ray Tracing + Shape Nearest Neighbor (PRP + NN): For a simple baseline that uses both unlabeled 3D shapes and 2D weak supervision, we propose a nearest neighbor retrieval of the unlabeled 3D shapes with PRP. We first use the PRP network to generate prediction and retrieve the nearest neighbor within the unlabeled 3D shapes. This method improves prediction accuracy if there is a similar shape among the unlabeled 3D shapes and the prediction from the PRP network is accurate.

Full 3D Supervision (F3D): Finally, we provide the results from full 3D supervision [5] as reference. The networks are trained with 3D supervision (3D shapes) on the same network architecture as in Sec. A.1 without adversarial discriminator. This experiment provides an upper bound performance for our WS-GAN since 2D projections only provide partial information of the 3D shapes.

A.3. Single-view reconst. on IKEA dataset[26]

In order to compare our work with the other recent supervised 3D reconstruction methods [14, 50], we tested our network on IKEA dataset. Similar to other works, we trained a single network on ShapeNet renderings of the furniture merged with random background from PASCAL [10]. Following the convention of [14, 50], we evaluated the reconstruction on ground-truth model aligned over permutations, flips, and translational alignments (up to 10%). Please note that all of the other baselines require full 3D supervision that is meant to provide upper bound performance over WS-GAN. The quantitative results can be found in Tab. A1.

Method	Chair	Desk	Sofa	Table	Mean
AlexNet-fc8[14]	20.4	19.7	38.8	16.0	23.7
AlexNet-conf4[14]	31.4	26.6	69.3	19.1	37.1
T-L Network[14]	32.9	25.8	71.7	23.3	39.6
3D-VAE-GAN[50]	42.6	34.8	79.8	33.1	48.8
WS-GAN	32.0	28.6	55.7	29.0	37.0

Table A1. Per-class real image 3D reconstruction Average Precision(AP) percentage on IKEA dataset[26]. Please note that all of the other baselines require full 3D supervision that are meant to provide upper bound performance over WS-GAN.

A.4. ObjectNet3D reconst. quantitative result

As shown in the main paper, we demonstrated a single-view real image reconstruction trained on ObjectNet3D[52] dataset. We quantitatively evaluated intersection-over-union(IOU) on the reconstruction results as shown in Table A2. The numbers indicate that WS-GAN is capable of learning 3D reconstruction beyond the issue of ill-conditioned visual hull reconstruction from a single-view mask. Please also note that voxel carving, unlike WS-GAN, requires camera parameters at test time.

	sofa	chair	bench	car	airplane
Voxel Carving	0.304	0.177	0.146	0.481	0.151
WS-GAN	**0.423**	**0.380**	**0.380**	**0.649**	**0.322**

Table A2. Per-class real image 3D reconstruction intersection-over-union(IOU) percentage on ObjectNet3D.

A.5. Multi-view synthetic images reconstruction

In Figure A2, we visualized more qualitative reconstruction results on ShapeNet [4] dataset. In order to visualize the strength and the weakness of WS-GAN, we presented both successful and less-successful reconstruction results. In general, as discussed in the main paper, WS-GAN reconstructed a reasonable 3D shape from a small number of silhouettes and viewpoints. However, WS-GAN had some difficulty reconstructing exotic shapes which might not be in the unlabeled shape repository given to the discriminator.

A.6. Single real image reconstruction

In Figure A3, we visualized more qualitative reconstruction results on ObjectNet3D[52] dataset. We observed that WS-GAN can learn to reconstruct a reasonable 3D shape from a single mask supervision.

A.7. Multi-view real image reconstruction

In Figure A4, we visualized more qualitative reconstruction results on Stanford Online Product Dataset [43]. As explained in the main paper, we trained the network on the ShapeNet [4] dataset with random background images from PASCAL [10] to make the network robust to the background noise. Since the domain of the train and test data are different, the reconstruction quality may not be as good as other experiments. However, our network shows reasonable 3D reconstruction results.

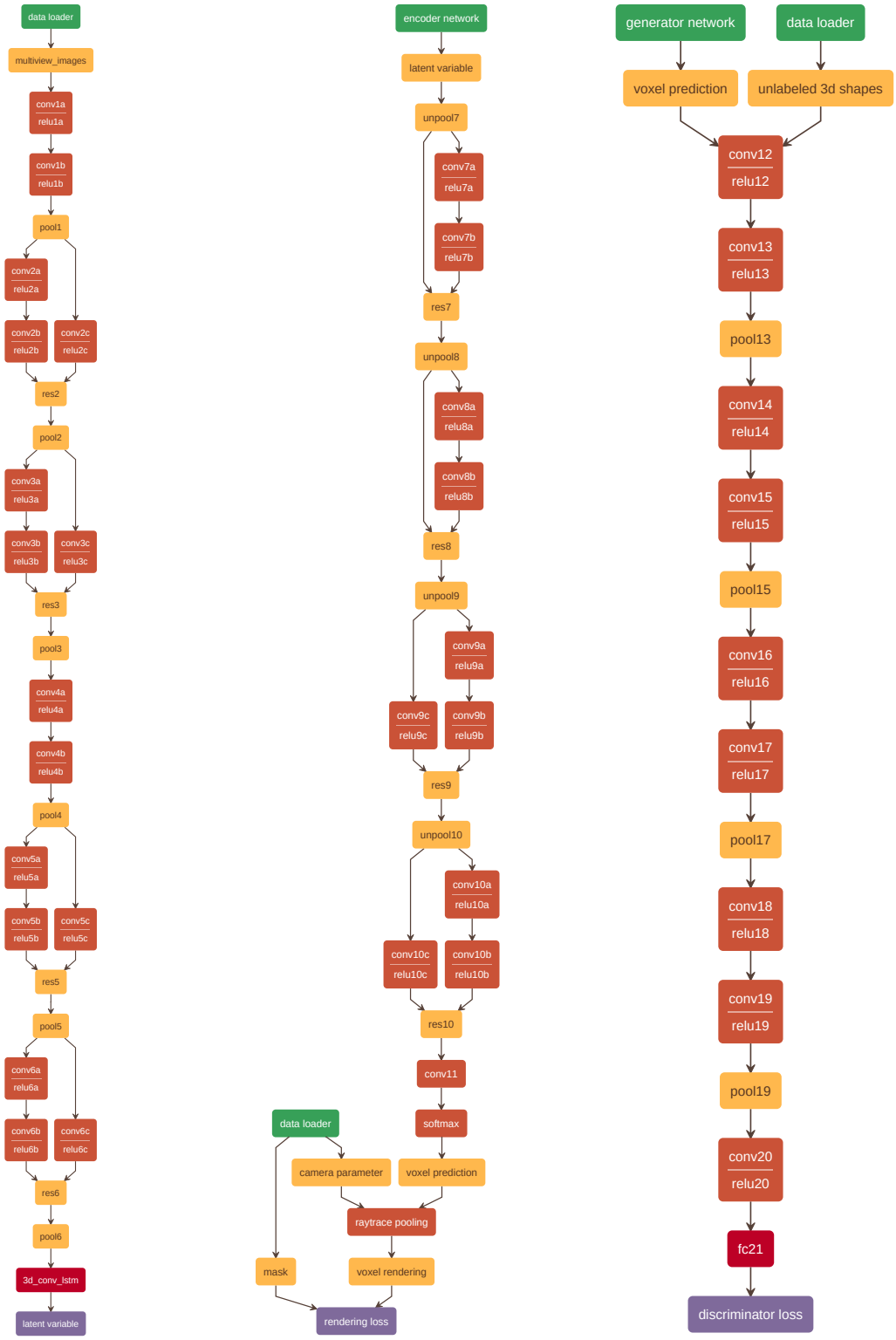
A.8. Representation analysis

We present more representation analysis results similar to the results in the main paper Sec. 5.6. In Figure A5, we linearly interpolate the latent variables of two images inter-and intra-class. This shows that the latent space that the encoder learned is the smooth space over the 3D shapes. In Figure A6, we add and subtract the latent variables of different images to modify the generated voxels with semantic context. Both experiments hint that the latent variable of WS-GAN has a meaningful semantic expressiveness that allows us to manipulate 3D shapes semantically.

A.9. Computation Time

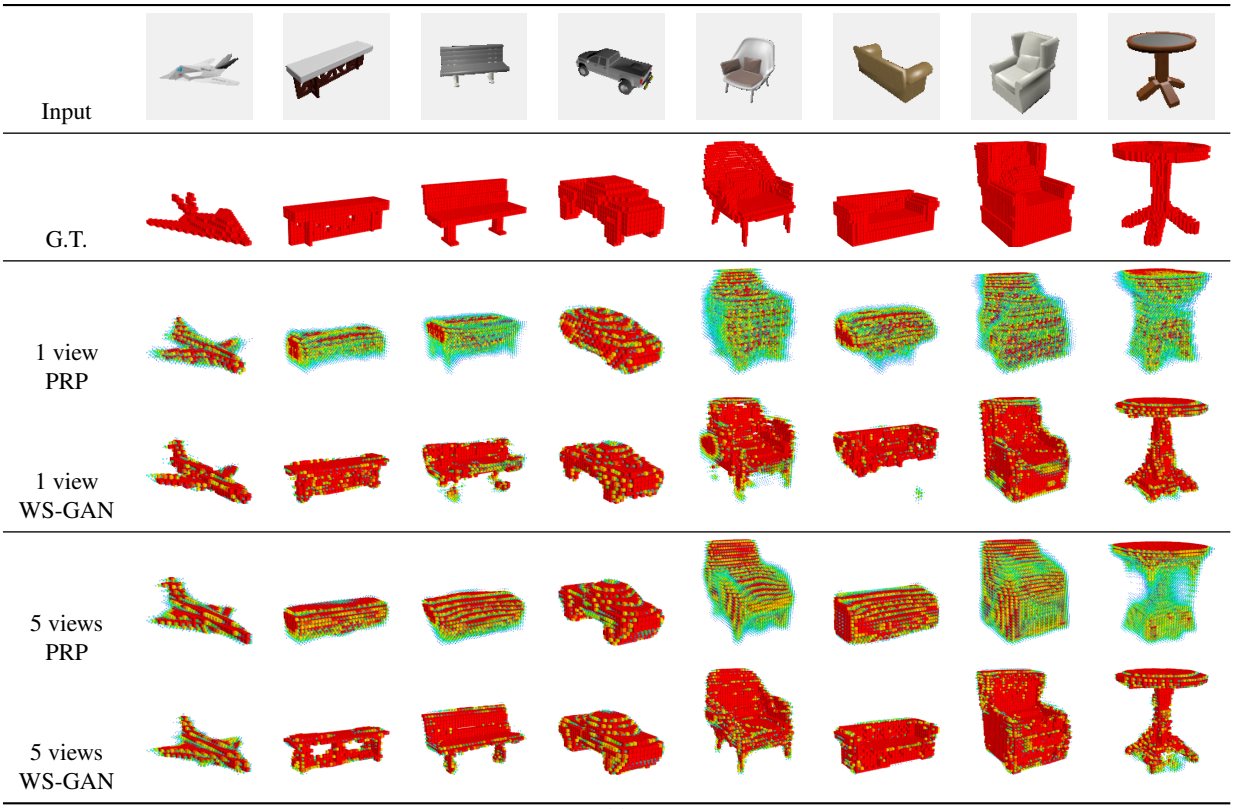
We evaluated computation time of all methods in our experiments. All experiments are on NVIDIA Titan X with batch size 8 and 5 views. Please note that at test time, we only evaluate encoder E and generator G, thus, the computation time is the same for PRP and WS-GAN.

Method	Voxel Carving	PRP train	WS-GAN train	WS-GAN test
Time(s)	0.115	3.57	5.16	0.268

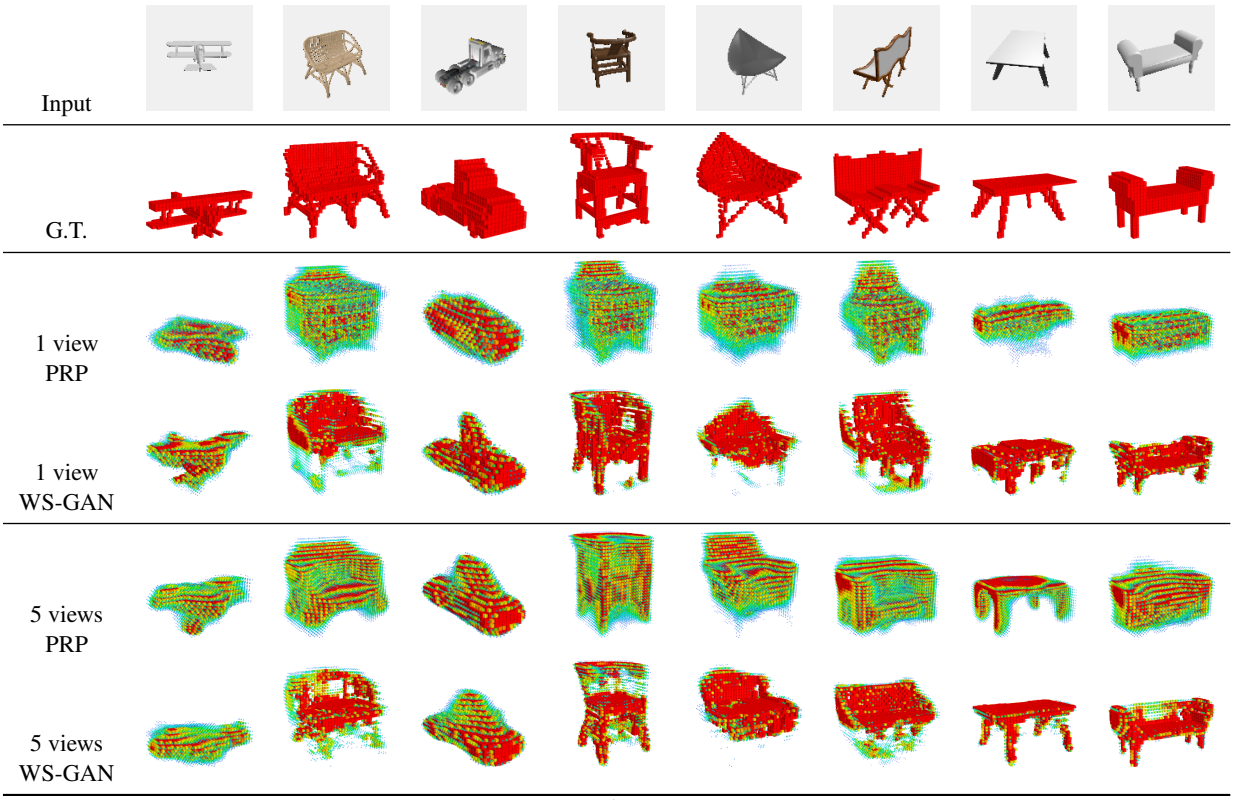


(a) encoder network (b) generator network (c) discriminator network

Figure A1. Detailed network structure of WS-GAN. Please note that all of these components are connected as a single network in our implementation. We split the figure into three for better visualization.



(a)



(b)

Figure A2. (a) Successful (b) less-successful qualitative results of single- or multi-view synthetic image reconstructions on ShapeNet dataset. This result hints that our WS-GAN is learning high-quality reconstruction including concavity from a small number of views of mask supervision. Please check the main paper for details of our visualization method.

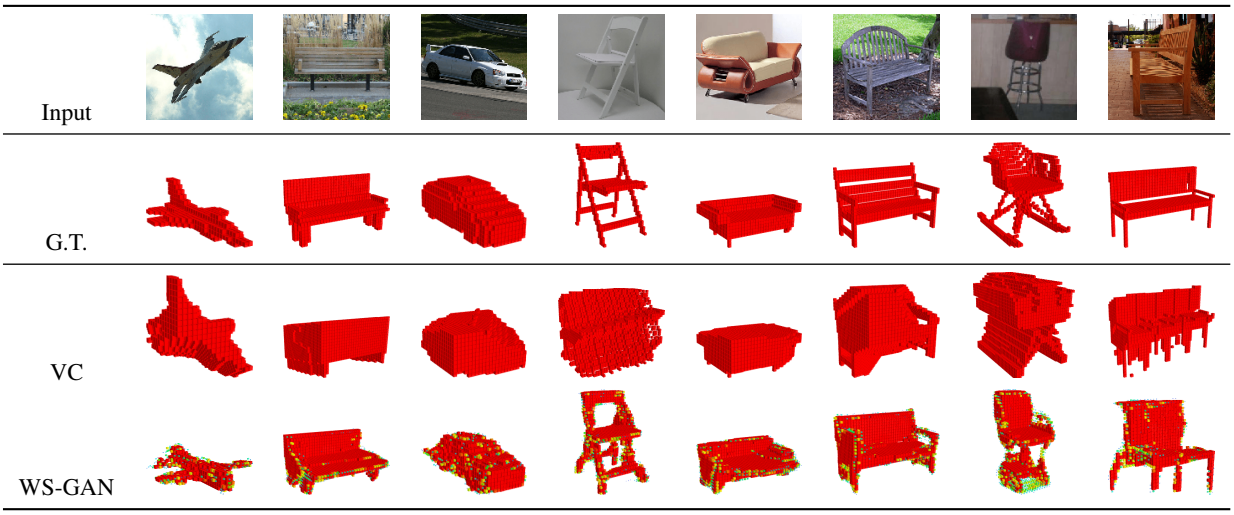


Figure A3. Qualitative results of real image reconstructions on ObjectNet3D. The results hints that our network successfully carved out concavity, which cannot be learned from mask supervision. Please note that voxel carving requires camera parameter at test time while ours does not.

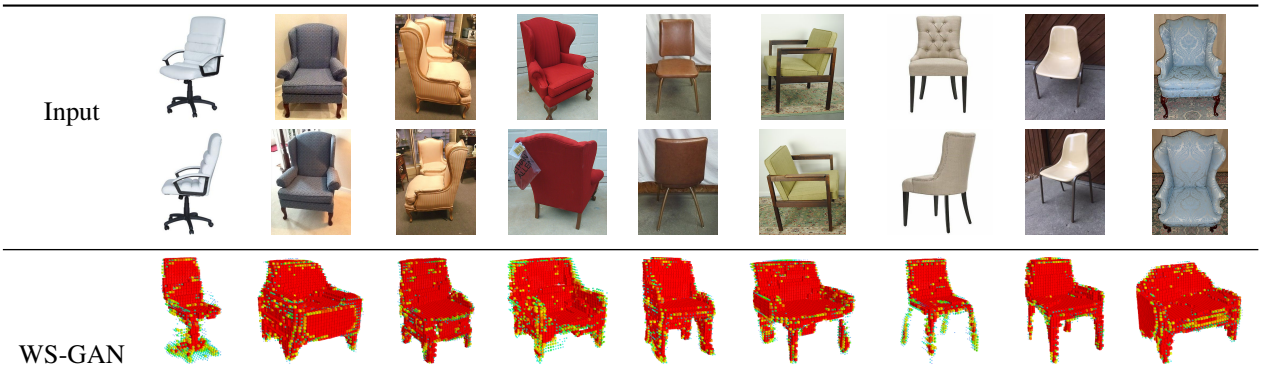


Figure A4. Qualitative results of multi-view real image reconstructions on Stanford Online Product dataset [43]. Our network successfully reconstructed real images coordinating multi-view information. Please note that the domain of training is different from that of test, which makes the reconstruction more challenging.

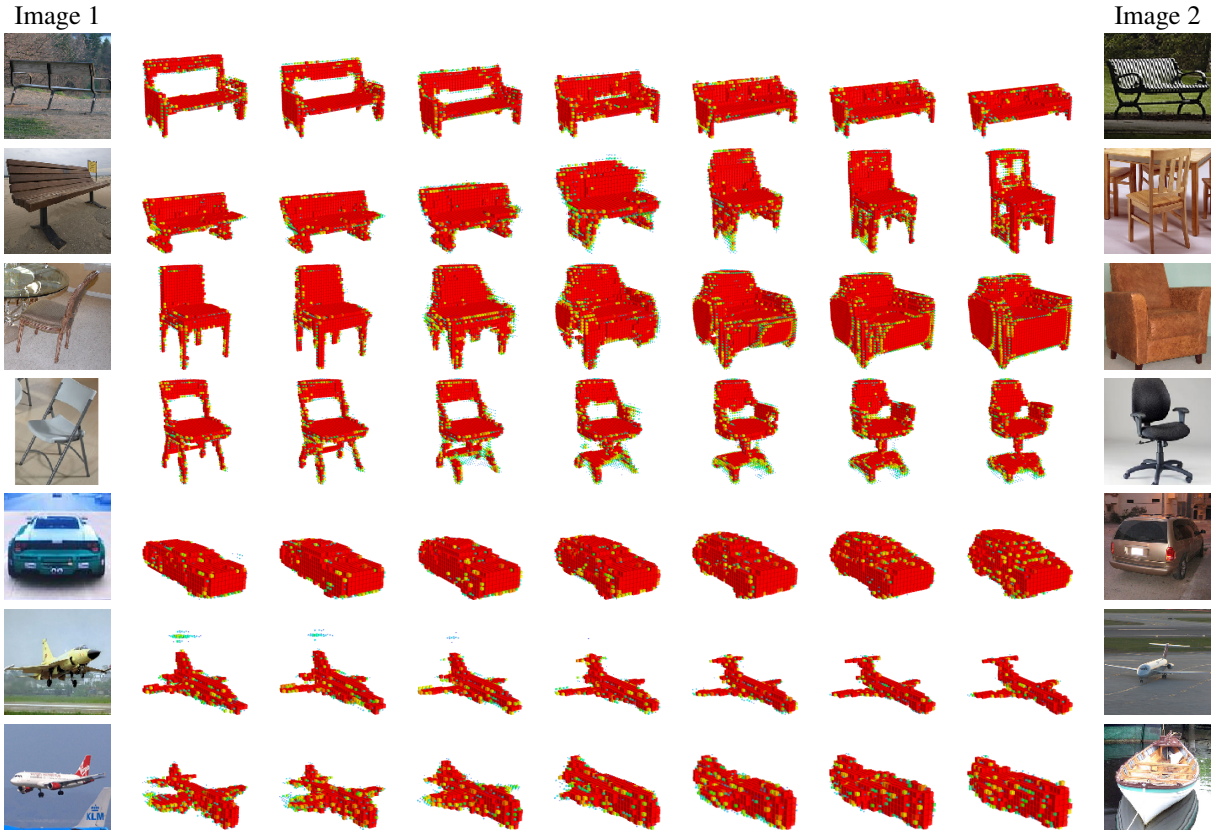


Figure A5. Linear interpolation of latent variable z. We observed the smooth transition of objects inter-and intra-class. Interestingly, semantic properties of the object, such as the length of the airplane wings and the size of the hole in the back of the chair smoothly transitioned. This result hints that our network generalized such semantic properties in the latent variable z.

Figure A6. Arithmetics on latent variable z of different images. By subtracting latent variables of similar chairs with different properties, we extracted the feature which represents such property. We applied the feature to two other chairs to demonstrate that this is a generic and replicable representation.