GAL: Geometric Adversarial Loss for Single-View 3D-Object Reconstruction

Li Jiang[1], Shaoshuai Shi[1], Xiaojuan Qi[1], and Jiaya Jia[1,2]

[1] The Chinese University of Hong Kong
[2] Tencent YouTu Lab
{lijiang, xjqi, leojia}@cse.cuhk.edu.hk ssshi@ee.cuhk.edu.hk

Abstract. In this paper, we present a framework for reconstructing a point-based 3D model of an object from a single-view image. We found distance metrics, like Chamfer distance, were used in previous work to measure the difference of two point sets and serve as the loss function in point-based reconstruction. However, such point-point loss does not constrain the 3D model from a global perspective. We propose adding geometric adversarial loss (GAL). It is composed of two terms where the geometric loss ensures consistent shape of reconstructed 3D models close to ground-truth from different viewpoints, and the conditional adversarial loss generates a semantically-meaningful point cloud. GAL benefits predicting the obscured part of objects and maintaining geometric structure of the predicted 3D model. Both the qualitative results and quantitative analysis manifest the generality and suitability of our method.

Keywords: 3D reconstruction · adversarial loss · geometric consistency · point cloud · 3D neural network

1 Introduction

Single-view 3D object reconstruction is a fundamental task in computer vision with various applications in robotics, CAD, virtual reality and augmented reality. Recently, data-driven 3D object reconstruction attracts much attention [3, 4, 7] with the availability of large-scale ShapeNet dataset [2] and advent of deep convolutional neural networks.

Previous approaches [3, 4, 7, 21] adopted two types of representations for 3D objects. The first is voxel-based representation that requires the network to directly predict the occupancy of each voxel [3, 7, 21]. Albeit easy to integrate into deep neural networks, voxel-based representation suffers from efficiency and memory issues, especially in high-resolution prediction. To address these issues, Fan *et al.* [4] proposed point-based representation, in which the object is composed of discrete points. In this paper, we design our system based on point-based representation considering its scalability and flexibility.

Along the line of forming point-based representation, researchers focused on designing loss functions to measure the distance between prediction point set and ground-truth set. Chamfer distance and Earth Mover distance were used

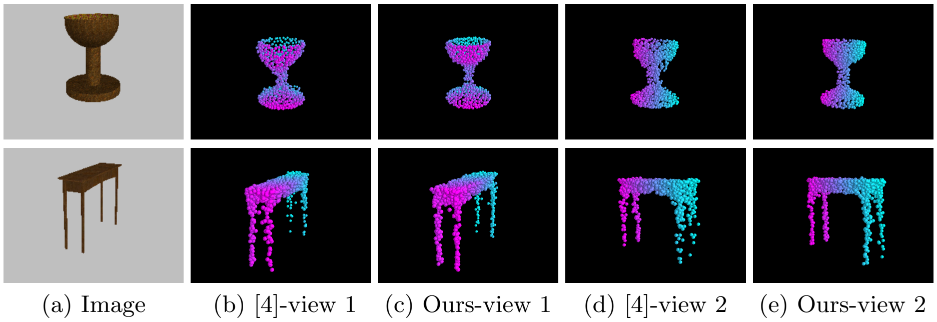


Fig. 1. Illustration of predictions. (a) Original image including the objects to be reconstructed. (b)&(d) Results of [4] when viewed in two different angles. (c)&(f) Our prediction results from corresponding views. Color represents the relative distance to the camera in (b)-(e).

in [4] to train the model. These functions penalize prediction deviating from the ground-truth location. The limitation is that there is no guarantee that the predicted points follow the geometric shape of objects. It is possible that the result does not lie in the manifold of the real 3D objects.

We address this problem in this paper and propose a new complementary loss function – *geometric adversarial loss (GAL)*. It regularizes prediction globally by enforcing the prediction to be consistent with the ground-truth among different 2D views and following the 3D semantics of point cloud.

GAL is composed of two important components, namely, geometric loss and conditional adversarial loss. Geometric loss lets the prediction in different views consistent with the ground truth. Regarding conditional adversarial loss, the conditional discriminator network combines a 2D CNN, to extract image semantic features, with PointNet [16], which extracts global features of the predicted/ground-truth point cloud. Features from the 2D CNN serve as a condition to enforce predicted 3D point cloud with respect to the semantic class of the input. In this regard, GAL regularizes predictions in a global perspective and thus can work in complement with previous CD [4] loss for better object reconstruction from a single image.

Fig. 1 preliminarily illustrates the reconstruction quality. When measured using chamfer distance, predictions by previous method [4] are similar to ours with just 0.5% difference. However, when viewed from different viewpoints, there come many noisy points as shown in Fig. 1(b)&(d) in the predicted point cloud produced by previous work. This is because the global 3D geometry is not respected, and only local point-to-point loss is adopted. With geometric adversarial loss (GAL) to regularize prediction globally, our method produces geometrically more reasonable results as shown in Fig. 1(c)&(e). Our main contribution is threefold.

- We propose a loss function, namely GAL, to geometrically regularize prediction from a global perspective.

- We extensively analyze contribution of different loss functions in generating 3D objects.
- Our method achieves better results both quantitatively and qualitatively in ShapeNet dataset.

2 Related Work

2.1 3D Reconstruction from Single Images

Traditional 3D reconstruction methods [10, 1, 13, 11, 8, 5] require multiple view correspondence. Recently, data-driven 3D reconstruction from single images [4, 3, 7, 21, 19] has gained more attention. Reconstructing 3D shapes from single images is ill-posed but desirable in real-world applications. Moreover, human actually have the ability to infer 3D shapes of objects given only a single view of it by using prior knowledge and visual experience of the 3D world. Previous work in this setting can be coarsely cast into two categories.

Voxel-based Reconstruction One stream of research focuses on voxel-based representation [3, 7, 21]. Choy *et al.* [3] proposed applying 2D convolutional neural networks to encode prior knowledge about the shape into a vector representation and then 3D convolutional neural network was used to decode the latent representation into 3D object shapes. Follow-up work [7] proposed the adversarial constraint to regularize predictions in the real manifold with a large amount of unlabeled realistic 3D shapes. Tulsiani *et al.* [20] adopted an unsupervised solution for 3D object reconstruction by jointly learning a pose estimation network and 3D object voxel prediction network with the multi-view consistency constraint.

Point Cloud Reconstruction Voxel-based representation may suffer from memory and computation issues when scaled to high resolutions. To address this issue, point cloud based representation for 3D reconstruction was introduced by Fan *et al.* [4]. Unordered point cloud is directly derived from a single image, which can encode more details of 3D shape. The end-to-end framework directly regresses point location. Chamfer distance is adopted to measure the difference between predicted point cloud and ground truth. We follow this line of research. Yet we make our contribution on a new differentiable multi-view geometric loss to measure results from different viewpoints, which is complementary to chamfer distance. We also use conditional adversarial loss as a manifold regularizer to make the predicted point cloud more reasonable and realistic.

2.2 Point Cloud Feature Extraction

Point cloud feature extraction is a challenging problem since points lie in a non-regular space and cannot be processed easily with common CNNs. Qi *et al.* [16] proposed PointNet to extract unordered point representation by using multilayer perceptron and global pooling. Transformer network is incorporated to learn robust transformation invariant features. PointNet is a simple and yet elegant framework to extract point features. As a follow-up work, PointNet++ was proposed in [17] to integrate global and local representations with much increased computation cost. In our work, we adopt pointNet as our feature extractor for predicted and ground truth point clouds.

2.3 Generative Adversarial Networks

There is a large body of work for generative adversarial networks [6, 22, 12, 14, 9] to create 2D images by regularizing prediction in the manifold of the target space. Generative adversarial networks were used in reconstructing 3D models from single-view images in [7, 21]. Gwak *et al.* [7] better utilized unlabeled data for 3D voxel based reconstruction. Yang *et al.* [21] reconstructed 3D object voxels from single depth images. They show promising results in a simpler setting since one view of the 3D model is given with accurate 3D position. Different from these approaches, we design a conditional adversarial network for 3D point cloud based reconstruction to enforce prediction in the same semantic space under the condition of using single-view images.

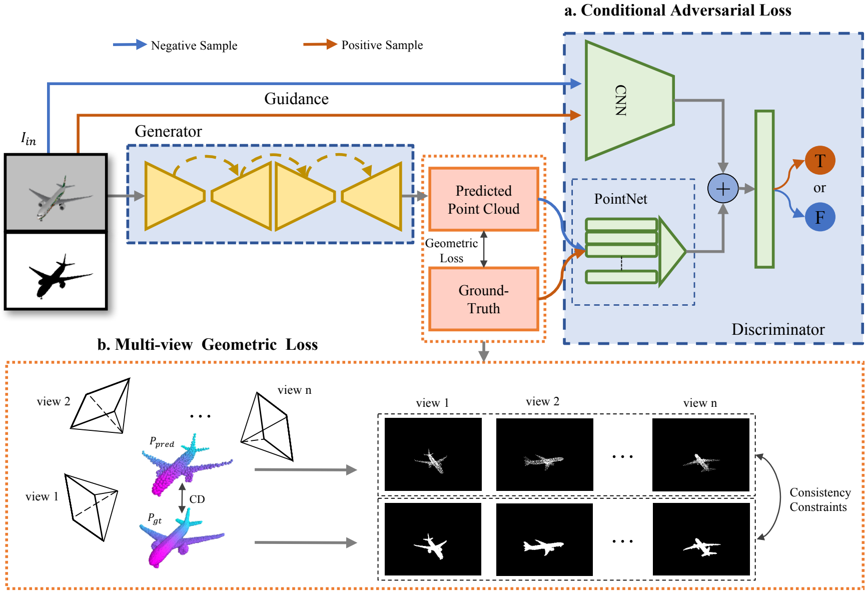


Fig. 2. Overview of our framework. The whole network consists of two parts: a generator network taking a single image as input and producing a point cloud modeling the 3D object, and a discriminator for judging the ground-truth and generated model conditioned on the input image. Our proposed geometric adversarial loss (GAL) is composed of conditional adversarial loss (a) and multi-view geometric loss (b).

3 Method Overview

Our approach produces 3D point cloud from a single-view image. The network architecture is shown in Fig. 2.

In the following, I_{in} denotes the input RGB image, and P_{gt} denotes the ground-truth point cloud. As illustrated in Fig. 2, the framework consists of two networks, i.e., generator network (G) and conditional discriminator network (D). G is the same as the one used in [4] composed of several encoder-decoder hourglass [15] modules and a fully connected branch to produce point locations. It is responsible for producing point locations that map input image I_{in} to its corresponding point cloud P_{pred}. Since it is not our major contribution, we refer readers to the supplementary material for more details.

The other component – conditional discriminator (D) (Fig. 2) – contains a PointNet [16] to extract features of the generated and ground-truth point clouds, and a CNN takes I_{in} as input to extract semantic features of the object. The extracted features are combined together as the final representation. The goal is to distinguish between the generated 3D prediction and the real 3D object.

Built upon the above network architecture, our loss function GAL regularizes the prediction globally to enforce it to follow the 3D geometry. GAL is composed of two components as shown in Fig. 2, i.e., multi-view geometric loss detailed in Section 4.1 and conditional adversarial loss detailed in Section 4.2. They work in synergy with the point-to-point chamfer-distance-based loss function [4] for both global and local regularization.

4 GAL: Geometric Adversarial Loss

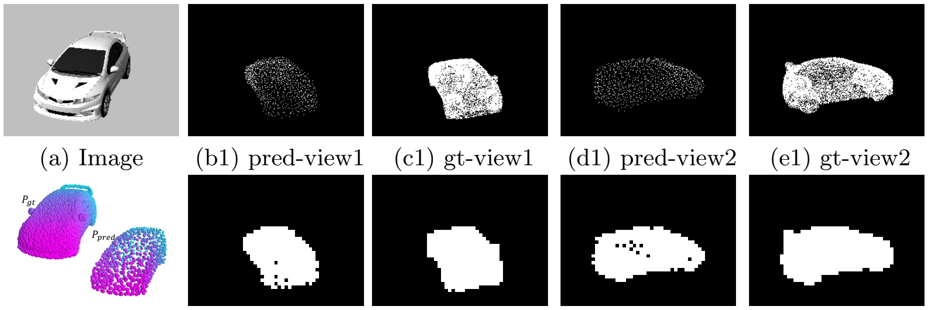
4.1 Multi-view Geometric Loss

Human can naturally figure out the shape of an object even if only one view is available. It is because of prior knowledge and knowing the overall shape of the objects. In this section, we add multi-view geometric constraints to inject such prior in neural networks. Multi-view geometric loss shown in Fig. 2 measures the inconsistency of geometric shapes between the predicted points P_{pred} and ground-truth P_{gt} in different views.

We first normalize the point clouds to be centered at the origin of the world coordinate. The numbers of points in P_{gt} and P_{pred} are respectively denoted as n_{gt} and n_p. n_p is pre-assigned to 1024 following [4]. n_{gt} is generally much larger than n_p.

To measure multi-view geometric inconsistency between P_{gt} and P_{pred}, we synthesize an image for each view given the point set and view parameters, and then compare each pair of images synthesized from P_{gt} and P_{pred}. Two examples are shown in Fig. 3(b1)-(e1).

To project the 3D point cloud to an image, we first transform point p_w with 3D world coordinate $p_w = (x_w, y_w, z_w)$ to camera coordinates $p_c = (x_c, y_c, z_c)$ as Eq. (1). R and d represent the rotation and translation parameters of the



(a) Image (b1) pred-view1 (c1) gt-view1 (d1) pred-view2 (e1) gt-view2

(f) Point cloud (b2) pred-view1 (c2) gt-view1 (d2) pred-view2 (e2) gt-view2

Fig. 3. (a) is the original image. (b1)&(d1) show the high resolution 2D projection of predicted point cloud in two different views. (c1)&(e1) show the high resolution 2D projection of the ground-truth point cloud. (b2)-(e2) show the corresponding low resolution results. (f) shows the ground-truth and predicted point cloud.

camera regarding the world coordinate. The rotation angles over $\{x, y, z\}$-axis are randomly sampled from $[0, 2\pi)$. Finally, point p_w is projected to the camera plane with function f as

$$p_c = Rp_w + d, \; f(p_w|K) = Kp_c, \tag{1}$$

where K is the camera intrinsic matrix.

We set the intrinsic parameters of our view camera as Eq. (2) to guarantee that the object is completely included in the image plane and the projected region occupies the image as much as possible.

$$u_0 = 0.5h, \; v_0 = 0.5w, \; f_u = f_v = \frac{0.5 \min(\{z_c\}) \min(h, w)}{\max(\{x_c\} \bigcup \{y_c\})} \tag{2}$$

where h and w are the height and width of the projected image.

Then, the projected images of ground-truth and predicted point cloud with size (h, w) could be respectively formulated as

$$I_{gt}^{h,w}(p) = \begin{cases} 1, & \text{if } p \in f(P_{gt}) \\ 0, & \text{otherwise} \end{cases}, \; I_{pred}^{h,w}(p) = \begin{cases} 1, & \text{if } p \in f(P_{pred}) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where p indexes over all the pixels of the projected image.

The synthesized views (Fig. 3) are with different densities in high resolutions. The projection images from ground-truth shown in Fig. 3(c1)&(e1) is much denser than our corresponding prediction shown in Fig. 3(b1)&(d1). To resolve the above discrepancy, multi-view geometric consistency loss is added in multiple resolutions detailed in the following.

High Resolution Mode In high resolution mode, we set h and w to large values denoted by h_1 and w_1 respectively. Images projected in this mode could contain details of the object as shown in Fig. 3(b1)-(e1). However, with the large difference between point amounts in P_{gt} and P_{pred}, the image projected from P_{pred} has less non-zero pixels than image projected from P_{gt}. Thus, calculating the L2 distance of the two images directly is not feasible. We define the high-resolution consistency loss for a single view v as

$$\mathcal{L}_v^{high} = \sum_p \mathbb{1}(I_{pred}^{h_1,w_1}(p) > 0) \| I_{pred}^{h_1,w_1}(p) - \max_{q \in N(p)} I_{gt}^{h_1,w_1}(q) \|_2^2, \tag{4}$$

where p indexes pixel coordinates, $N(p)$ is the $n \times n$ block centered at p, and $\mathbb{1}(.)$ is an indicator function set to 1 when the condition is satisfied. Since the predicted point cloud is sparser than the ground-truth, we only use the non-zero pixels in the predicted image to measure the inconsistency. For each non-zero pixel in I_{pred}, we find the corresponding position in I_{gt} and search its neighbors for non-zero pixels to reduce the influence of projection errors.

Low Resolution Mode In the high-resolution mode, we only check whether the non-zero pixels in I_{pred} appear in I_{gt}. Note that the constraint needs to be bidirectional. We make I_{pred} the same density as I_{gt} by setting h and w to small values h_2 and w_2. Low-resolution projection images are shown in Fig. 3(b2)-(e2). Although details are lost in the low resolution, rough shape is still visible and can be used to check the consistency. Thus, we define the low-resolution consistency loss for a single view v as

$$\mathcal{L}_v^{low} = \sum_p \| I_{pred}^{h_2,w_2}(p) - I_{gt}^{h_2,w_2}(p) \|_2^2, \tag{5}$$

Where $I_{pred}^{h_2,w_2}$ and $I_{gt}^{h_2,w_2}$ represent the low resolution projection images and h_2 and w_2 are the corresponding height and width. The low-resolution loss constrains that the shapes of ground-truth and predicted objects are similar, while the high-resolution loss ensures the details.

Total Multi-view Geometric Loss We denote v as the view index. The total multi-view geometric loss is defined as

$$\mathcal{L}_{mv} = \sum_v (\mathcal{L}_v^{high} + \mathcal{L}_v^{low}). \tag{6}$$

The objective regularizes the geometric shape of predicted point cloud from different viewpoints.

4.2 Point-based Conditional Adversarial Loss

To generate a more plausible point cloud, we propose using a conditional adversarial loss to regularize the predicted 3D object points. The generated 3D model should be consistent with the semantic information provided by the image. We adopt PointNet [16] to extract the global feature of the predicted point cloud. Also, with the 2D semantic feature provided by the original image, the discriminator could better distinguish between the real 3D model and the generated fake one. Thus, the RGB image of the object is also fed into the discriminator. P_{pred} along with the corresponding I_{in} serve as a negative sample, while P_{gt} and I_{in} become positive when training the discriminator. During the course of training the generator, the conditional adversarial loss forces the generated point cloud to respect the semantics of the input image.

The CNN part of the discriminator is a pre-trained classification network to extract 2D semantic features, which are then concatenated with feature produced by PointNet [16] for identifying real and fake samples. We note that the point cloud from our prediction is sparser than ground-truth. Hence, we uniformly sample n_p points from ground-truth with a total of n_{gt} points.

Different from traditional GAN, which may be unstable and has low convergence rate, we apply LSGAN as our adversarial loss. LSGAN replaces logarithmic loss function with least-squared loss, which makes it easier for the generated data distribution to converge to the decision boundary. The conditional adversarial loss function is defined as

$$\begin{aligned}\mathcal{L}_{LSGAN}(D) &= \frac{1}{2}[\mathbb{E}_{P_{gt}\sim p(P_{gt})}(D(P_{gt}|I_{in})-1)^2+\mathbb{E}_{I_{in}\sim p(I_{in})}(D(G(I_{in})|I_{in})-0)^2] \\ \mathcal{L}_{LSGAN}(G) &= \frac{1}{2}[\mathbb{E}_{I_{in}\sim p(I_{in})}(D(G(I_{in})|I_{in})-1)^2]\end{aligned} \tag{7}$$

During the training process, G and D are optimized alternately. G minimizes $\mathcal{L}_{LSGAN}(G)$, which aims to generate a point cloud similar to the real model, while D minimizes $\mathcal{L}_{LSGAN}(D)$ to discriminate between real and predicted point sets. In the testing process, only the well-trained generator needs to be used to reconstruct a point cloud model from a single-view image.

5 Total Objective

To better generate a 3D point cloud model from a single-view image, we combine the conditional adversarial loss and the geometric consistency loss as GAL for global regularization. We also follow the distance metric in [4] to use Chamfer distance to measure the point-to-point similarity of two point sets as a local constraint. Chamfer distance loss is defined as

$$\mathcal{L}_{cd}(I_{in},P_{gt}|G)=\frac{1}{n_{gt}}\sum_{p\in P_{gt}}\min_{q\in G(I_{in})}\|p-q\|_2^2+\frac{1}{n_p}\sum_{p\in G(I_{in})}\min_{q\in P_{gt}}\|p-q\|_2^2. \tag{8}$$

With global GAL and point-to-point distance constraint, the total objective becomes

$$\begin{aligned}G^* &= arg\min_G[\mathcal{L}_{LSGAN}(G)+\lambda_1\mathcal{L}_{mv}+\lambda_2\mathcal{L}_{cd}] \\ D^* &= arg\min_D\mathcal{L}_{LSGAN}(D)\end{aligned} \tag{9}$$

where λ_1 and λ_2 control the ratio of different losses.

The generator is responsible for fooling the discriminator, and reconstructing a 3D point set approximating the ground-truth. The adversarial part ensures the reconstructed 3D object to be reasonable with respect to the semantics of the original image. Multi-view geometric consistency loss makes the predicted point cloud a valid prediction when viewed in different directions.

6 Experiments

We perform our experiments on the ShapeNet dataset [2], which has a large collection of textured CAD models. Our detailed network architecture and implementation strategies are the following.

Generator Architecture Our generator G is built upon the network structure in [4], which takes a 192×256 image as input and consists of a convolution branch producing 768 points and a fully connected branch producing 256 points, resulting in total 1024 points.

Discriminator Architecture Our discriminator D contains a CNN part to extract semantic features from the input image and a PointNet part to extract features from point cloud as shown in Fig. 2. The backbone of the CNN part is VGG16 [18]. A fully connected layer is added after the $fc8$ layer to reduce the feature dimension to 40.

The major building block in PointNet is multi-layer perceptron (MLP) and global pooling as in [16]. The MLP utilized on points contains 5 hidden layers with layer sizes $(64, 64, 64, 128, 1024)$. The MLP after max pooling layer consists of 3 layers with sizes $(512, 256, 40)$. The features from CNN and PointNet are concatenated together for final discrimination.

Implementation Details The whole network is trained in an end-to-end fashion using ADAM optimizer with batch size 32. The view number for multi-view geometric loss is set to 7, which is determined by experimenting with different view numbers and selecting the one that gives the best performance. h_1, w_1, h_2, and w_2 are set to 192, 256, 48, and 64 respectively. The block size for neighborhood searching in high resolution mode is set to 3×3.

6.1 Ablation Studies

Evaluation Metric We evaluate the predicted point clouds of different methods using three metrics: point cloud based Chamfer Distance (CD), voxel based Intersection over Union (IoU) and 2D projection IoU. CD measures the distance between ground-truth point set and predicted one. The definition of CD is in Section 5. The lower CD value represents the better reconstructed results.

To compute IoU of two point sets, each point set will be voxelized by distributing points into 32 × 32 × 32 grids. We treat each point as a 1 × 1 × 1 grid centered at this point, namely point grid. For each voxel, we consider the maximum intersecting volume ratio of each point grid and this voxel as the occupancy probability. It is then translated into two-value form by a threshold t. The calculation formula of IoU is

$$IoU = \frac{\sum_i \mathbb{1}[V_{gt}(i)V_p(i) > 0]}{\sum_i \mathbb{1}[V_{gt}(i) + V_p(i) > 0]}, \tag{10}$$

where i indexes all voxels, $\mathbb{1}$ is an indicator function, V_{gt} and V_p are respectively the voxel-based ground-truth and voxel-based prediction. The higher IoU value indicates more precise point cloud prediction.

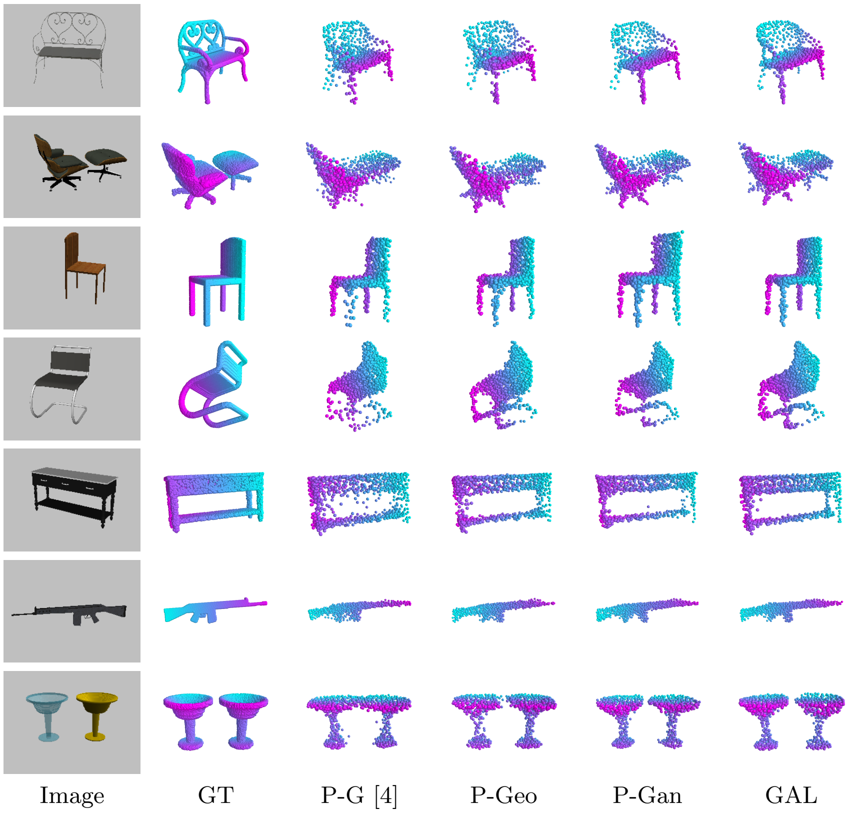


Fig. 4. Qualitative results of single image 3D reconstruction from different methods. For the same object, all the point clouds are visualized from the same viewpoint.

To better evaluate our generated point cloud, we propose a new projected view evaluation metric, i.e. 2D projection IoU, where we project the point clouds into images from different views, and then compute 2D intersection over union (IoU) between the ground-truth projected images and the reconstruction projected images. Here we use three views, namely top view, front view and left view, to evaluate the shape of generated point cloud comprehensively. And three resolutions are adopted, which are 192 × 256, 96 × 128, 48 × 64 respectively.

Comparison among Different Methods To thoroughly investigate our proposed GAL loss, we consider the following settings for ablation studies.

- *PointSetGeneration*(P-G) [4], which is a point-form single image 3D object reconstruction method. We directly use the model trained by the author-released code as our baseline.
- *PointGeo*(P-Geo), which combines the geometric loss proposed in Section 4.1 with our baseline to evaluate the effectiveness of geometric loss.
- *PointGan*(P-Gan), which combines the point-based conditional adversarial loss with our baseline to evaluate the effectiveness of adversarial loss.
- *PointGAL*(GAL), which is the complete framework as shown in Fig. 2 to evaluate the effectiveness of our proposed GAL loss.

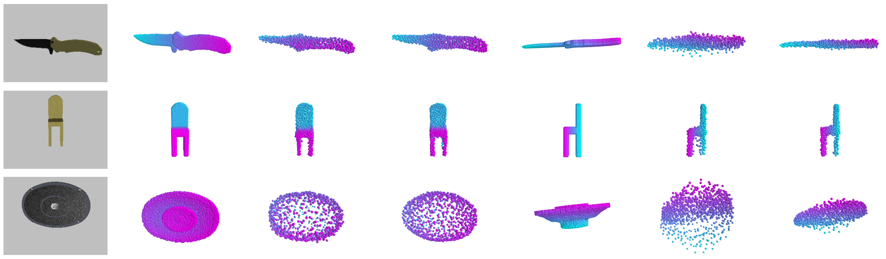
Table 1. Ablative results over different loss functions.

	CD$\times 10^{-4}$ (lower is better)				IoU% (higher is better)			
	P-G	P-Geo	P-Gan	GAL	P-G	P-Geo	P-Gan	GAL
couch	39.15	37.59	37.88	**34.35**	71.71	72.08	72.37	**73.87**
cabinet	22.94	23.08	**22.27**	22.72	77.61	77.33	**77.79**	77.22
bench	30.77	29.55	29.06	**27.24**	67.90	68.65	69.44	**70.85**
chair	37.54	36.72	36.51	**33.59**	66.81	67.81	68.35	**70.02**
monitor	14.65	15.06	**13.76**	14.93	78.99	79.40	79.92	**80.39**
firearm	44.23	44.16	**41.66**	42.33	66.76	68.62	69.86	**71.50**
speaker	44.10	43.08	47.24	**41.99**	67.68	68.20	68.44	**69.81**
lamp	39.19	39.18	42.39	**38.25**	66.48	67.50	68.56	**69.98**
cellphone	31.81	32.04	33.30	**28.29**	75.72	75.98	75.86	**77.30**
plane	80.20	77.01	78.10	**76.34**	65.20	66.86	66.85	**68.53**
table	32.67	31.00	30.10	**28.30**	67.93	69.08	69.85	**71.38**
car	40.51	38.61	39.10	**36.06**	72.05	72.81	72.51	**73.68**
watercraft	34.33	34.63	34.29	**33.01**	66.08	66.03	66.57	**67.50**
mean	40.90	39.62	39.79	**37.26**	68.07	69.10	69.64	**71.16**

Table 1 shows quantitative results regarding CD and IoU for 13 major categories following the setting of [4]. The statistics show that our PointGeo and PointGan models outperform the baseline method [4] in terms of both CD and IoU metrics. The final GAL model can further boost the performance and outperforms the baseline by a large margin. As shown in Table 2, GAL consistently improves 2D projection IoU in all viewpoints, which demonstrates the effectiveness of constraining geometric shape across different viewpoints.

Table 2. 2D projection IoU comparison. The images are projected with three resolutions for three different view points.

	Resolution 192x256				Resolution 96x128				Resolution 48x64			
	P-G	P-Geo	P-Gan	GAL	P-G	P-Geo	P-Gan	GAL	P-G	P-Geo	P-Gan	GAL
Front view	0.328	0.333	0.334	**0.340**	0.601	0.611	0.613	**0.622**	0.773	0.780	0.782	**0.792**
Left view	0.325	0.330	0.330	**0.337**	0.586	0.594	0.594	**0.606**	0.750	0.757	0.758	**0.770**
Top view	0.343	0.346	0.349	**0.355**	0.652	0.657	0.663	**0.673**	0.823	0.829	0.832	**0.839**
Mean-IoU	0.332	0.337	0.338	**0.344**	0.613	0.621	0.623	**0.634**	0.782	0.789	0.791	**0.801**



(a) Image (b) GT-v1 (c) P-G-v1 (d) P-Geo-v1 (e) GT-v2 (f) P-G-v2 (g) P-Geo-v2

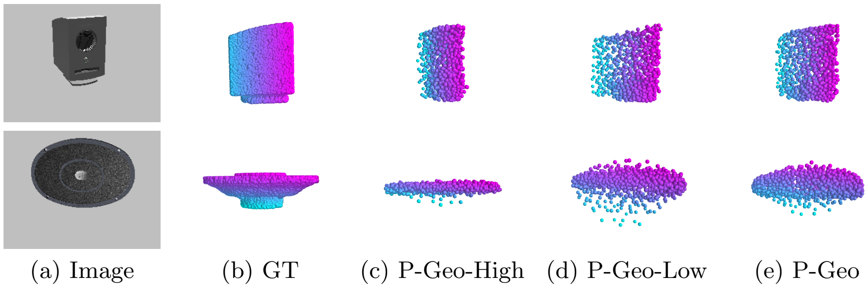
Fig. 5. Visualization of point clouds predicted by the baseline model (P-G) and our network with geometric loss (P-Geo) from two representative viewpoints. (b)-(d) are visualized from the viewpoint of the input image (v1), while (e)-(g) are synthesized from another view (v2).

Qualitative comparison is shown in Fig. 4. P-G [4] predicts less accurate structure where shape distortion arises (see the leg of furnitures and the connection between two objects). On the contrary, our method can handle these challenges and produce better results, since GAL penalizes inaccurate points from different views and regularizes prediction with semantic information from 2D input images.

Analysis of Multi-view Geometric Loss We analyze the importance of our multi-view geometric loss by checking the shape of the 3D models from different views. Fig. 5 shows two different views of the 3D model produced by the baseline model (P-G) and the baseline model with multi-view consistency loss (P-Geo).

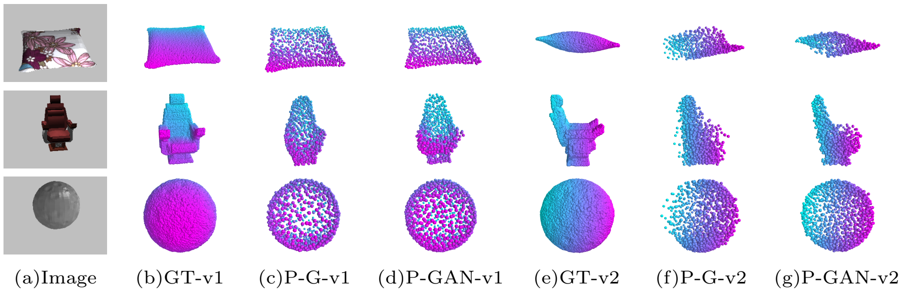
P-G result seems to be comparable (Fig. 5(c)) with ours shown in Fig. 5(d) when observed from the input image view angle. However, when the viewpoint changes, the generated 3D model of P-G (Fig. 5(f)) may not fit the geometry of the object. The predicted shape is much different from the real shape (Fig. 5(b)). In contrast, our reconstructed point cloud in Fig. 5(e) is still consistent with the ground-truth. When trained with multi-view geometric loss, the network penalizes incorrect geometric appearance from different views.

Analysis of Different Resolution Modes We have conducted the ablation study to analyze the effectiveness of different resolution modes. With only the



(a) Image (b) GT (c) P-Geo-High (d) P-Geo-Low (e) P-Geo

Fig. 6. Visualization of point clouds predicted in different resolution modes. P-Geo-High: P-Geo without low-resolution loss. P-Geo-Low: P-Geo without high-resolution loss.



(a)Image (b)GT-v1 (c)P-G-v1 (d)P-GAN-v1 (e)GT-v2 (f)P-G-v2 (g)P-GAN-v2

Fig. 7. P-G denotes our baseline model, P-GAN denotes the baseline model with conditional adversarial loss. Two different views are denoted by "v1" and "v2".

high-resolution geometric loss, the predicted points may lie inside the geometric shape of the object and do not cover the whole object as shown in Fig. 6(c). However, with only the low-resolution geometric loss, points may cover the whole object; but noisy points appear out of the shape as shown in Fig. 6(d). Combining the high and low-resolution loss, our trained model produces the best results as shown in Fig. 6(e).

Analysis of Point-based Conditional Adversarial Loss Our point-based conditional adversarial loss helps produce better semantically meaningful 3D object models.

Fig. 7 shows the pairwise comparison between the baseline model (P-G) and baseline model with conditional adversarial loss (P-GAN) from two different views. Without exploring the semantic information, the generated point clouds from P-G (Fig. 7(c)&(f)) seem contrived, while our results (Fig. 7(d)&(g)) look more natural from different views. For example, the chair generated by P-G cannot be recognized as a chair when observing from the side view (Fig. 7(f)), while our results have much better appearance seen from different directions.

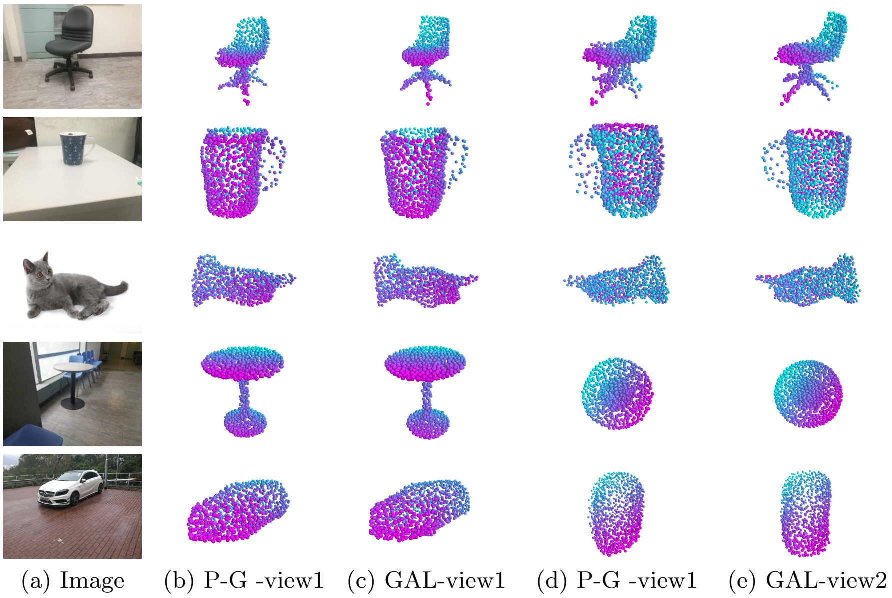


Fig. 8. Illustration of the real-world cases. (a) is the input image. (b) and (d) show results of P-G [4] from two different view angles. (c) and (f) show our prediction results from corresponding views.

6.2 Results on Real-world Objects

We also test the baseline and our GAL model on the real-world images. The images are manually annotated to get the mask of objects. The final results are shown in Fig. 8. Compared with the baseline method, the point clouds generated by our model capture more details. And in most cases, the geometric shape of our predicted point cloud seems to be more accurate in various views.

7 Conclusion

We have presented the geometric adversarial loss (GAL) to regularize single-view 3D object reconstruction from a global perspective. GAL includes two components, i.e. multi-view geometric loss and conditional adversarial loss. Multi-view geometric loss enforces the network to learn to reconstruct multiple-view valid 3D models. Conditional adversarial loss stimulates the system to reconstruct 3D object regarding semantic information in the original image. Results and analysis in the experiment section show that the model trained by our GAL achieves better performance on ShapeNet dataset than others. It can also generate precise point cloud from the real-world images. In the future, we plan to extend GAL to large-scale general reconstruction tasks.

References

1. Broadhurst, A., Drummond, T.W., Cipolla, R.: A probabilistic framework for space carving. In: ICCV (2001)
2. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv (2015)
3. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: ECCV (2016)
4. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In: CVPR (2017)
5. Fuentes-Pacheco, J., Ruiz-Ascencio, J., Rendón-Mancha, J.M.: Visual simultaneous localization and mapping: a survey. Artificial Intelligence Review (2015)
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS (2014)
7. Gwak, J., Choy, C.B., Chandraker, M., Garg, A., Savarese, S.: Weakly supervised 3d reconstruction with adversarial constraint (2017)
8. Häming, K., Peters, G.: The structure-from-motion reconstruction pipeline–a survey with focus on short image sequences. Kybernetika (2010)
9. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. arXiv (2017)
10. Laurentini, A.: The visual hull concept for silhouette-based image understanding. PAMI (1994)
11. Liu, S., Cooper, D.B.: Ray markov random fields for image-based 3d modeling: Model and efficient inference. In: CVPR (2010)
12. Lu, Y., Tai, Y.W., Tang, C.K.: Conditional cyclegan for attribute guided face image generation. arXiv (2017)
13. Matusik, W., Buehler, C., Raskar, R., Gortler, S.J., McMillan, L.: Image-based visual hulls. In: Proceedings of the 27th annual conference on Computer graphics and interactive techniques (2000)
14. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv (2014)
15. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV (2016)
16. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation (2017)
17. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: NIPS (2017)
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv (2014)
19. Tatarchenko, M., Dosovitskiy, A., Brox, T.: Multi-view 3d models from single images with a convolutional network. In: ECCV (2016)
20. Tulsiani, S., Efros, A.A., Malik, J.: Multi-view consistency as supervisory signal for learning shape and pose prediction. arXiv (2018)
21. Yang, B., Wen, H., Wang, S., Clark, R., Markham, A., Trigoni, N.: 3d object reconstruction from a single depth view with adversarial learning. arXiv (2017)
22. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv (2017)