Learning 3D Priors with Adversarial Novel View Generation

Aartika Rai
University of Massachusetts
Amherst
aartikarai@umass.edu

Abstract

We study the problem of learning single image 3D reconstruction for the scenario when no 3D supervision is available for training. Our approach relies on one or many images of each object and their corresponding viewpoints. The method we propose is a combination of multi-view consistency for learning 3D shape priors and generative adversarial networks (GANs) for generating images. Training a reconstructor using a single or a few views per object is a challenging task due the ambiguity in the shape from unobserved viewpoints. We show that our approach produces visually reasonable shapes even when the dataset consists of one view per object. We resolve the ambiguity by explicitly generating images for unobserved viewpoints by leveraging the ability of GANs to produce realistic high fidelity images. This image generation network is in-turn supervised by the projections of the produced shapes from these unobserved views. The key idea of our paper is a mutual supervision between these two components. While projecting the generated shape from another viewpoint produces a view that is geometrically consistent with the input images, the GAN produces images that fit within the distribution of real images.

1. Introduction

The task of predicting three dimensional shapes of objects from images or low dimensional probabilistic spaces, known as 3D reconstruction, is an active area of research. Generating 3D model of an object from a single or a few images has a wide range of applications in e-commerce, visualization, animation and architecture. Although it is an extremely ill-posed problem, humans can easily understand the underlying 3D structure of scenes and objects from their pictures. Using shading cues in the image to infer depth in conjunction with an intuition of what real world objects look like, humans are able to perform this wonderful feat with no effort. An interesting research question that arises from this observations is whether machines can be programmed to acquire this ability. For machines to be able to copy this ability they need to be able to perceive the 3D structure of the observed part as well as be able to reason about the unobserved viewpoints.

The focus of our work is to train a reconstructor using a small number of views of an object without explicit 3D supervision. Recent approaches [21, 16, 10] have utilized the principle of geometric consistency to learn 3D shape prediction using weaker signals such as images and pose information. Multi-view consistency is the principle that a common geometry, observed from different perspectives, can consistently explain multiple views of an instance. A straightforward application of this principle necessarily requires more than a single image per training object and increasing the number of views often leads to better results. Our proposed framework allows us to overcome this constraint and learn single-view shape prediction with data containing only a single view per instance. Therefore, our proposed method is novel in it's capability of utilizing the well tested principle of multi-view consistency in a more constrained set-up with single images and their viewpoints. When multiple views are available, however, our method offers diminishing returns and we show it's effectiveness with up to 4 views per object.

Deep convolutional neural networks have the capacity to produce shapes that are consistent with the provided ground truth view, but may have arbitrary shapes when viewed from other angles. An idea that has been studied before [5, 11] is to fix this issue by forcing the projections of the produced shape from other view points to fit the distribution of real images using an adversarial loss. Gadelha *et al*. [5] added a discriminator to directly differentiate the projections from the set of set the real images. While Kato and Harada [11] found it more effective to discriminate between projections from observed and unobserved viewpoints. Our approach follows the same idea but further decouples the 3D reconstructor from the adversarial component. We observe that this flexibility results in more stable training and better results in our experiments. We add a novel view generator to our model which provides explicit supervision to the 3D

shape generator about the object from unobserved viewpoints. We train this novel view generator with a combination of direct and adversarial supervision. The direct supervision is obtained by projecting the produced shape from the corresponding viewpoint. While the produced shape may be distorted from unobserved angles, this projection and the input image (being close to the projection from the original viewpoint) should still be geometrically consistent. We also add a discriminator which penalizes the novel view generator when the images do not match the real-image distribution. This loss is propagated to the 3D generator as well through the consistency loss between the projections and the generated images. Thus, our model is such that the two components i.e. the 3D reconstructor and the novel view generator are trained jointly and the output of one is used to supervise the other. The objective of these two components being mutually complimentary we find that this training regime produces promising results.

Our approach summarized in 1 learns shape and novel view prediction by enforcing geometric consistency between these predictions. Concretely, given one image of an object we predict a corresponding shape. In parallel, for the same image, we independently predict another image of the same object from a different perspective by randomly sampling a viewpoint. Then we enforce that the predicted shape should be 'consistent' with the novel view when observed from this sampled viewpoint.

Our formulation is fundamentally different from many recent works [18, 23, 20, 7, 22, 17] as we do not require any direct 3D supervision. This is motivated by the fact that it is much harder to procure a large set of 3D models as a human needs to design and assemble each model while images are cheap. Other works such as Yan *et al*. [21] and Tulsiani *et al*. [16] have also addressed this problem by proposing a networks which learns a prior on 3D shapes from multiple views of each object in the training set, and when the system is done learning it is able to give an estimate of the 3D shape of a novel object from only a single image. These methods use view-consistency to ensure that the projection of the 3D reconstruction from other views is consistent with the projected images from these viewpoints. However, our model is capable of learning a 3D prior given just a single image for each instance. In the absence of many views of an object, we use view-plausibility instead of view-consistency i.e. the projections of the 3D reconstructions should be plausible images of real-world objects.

2. Related Work

The task of predicting three dimensional shapes of objects from images or a low dimensional probabilistic space, known as 3D reconstruction, is a very active area of research. Recent works have explored different representations of the three dimensional world with voxels [19, 3, 21], point cloud [7, 4] and meshes [17, 9, 10] being the most dominant. While the voxel representation allows for a straightforward extension of image models to 3D, much of the recent works have focused on methods for point clouds and meshes as they can be scaled to many times higher resolution. In our work we use the voxel representation, however we note that the ideas are general and can be applied to point clouds.

A simple and popular approach for learning-based 3D reconstruction is to use 3D annotations. Wu *et al*. [19] proposed a 3D GAN architecture for unsupervisede generation of 3D models. MarrNet by Wu *et al*. [18] learns 3D shapes by first predicting 2.5 sketches. Zhang *et al*. [23] build on this work and generate 3D models for unseen classes by completing a spherical shape representation. Zhu *et al*. [20] learn to synthesize realistic novel view images by first constructing a 3D voxel representation. These works use voxels to represent 3D, but there has also been much work using point clouds and meshes as the 3D representation. AtlasNet [7] and FoldingNet [22] learn single view reconstruction of point clouds by representing 3D shapes as a collection of parametric surface elements. Pix2Mesh [17] represents 3D mesh in a graph-based convolutional neural network.

To reduce the cost of 3D annotation, view-based training has recently become an active research area. The key of training is to define a differentiable projection function. Perspective transformer net [21] reconstructs 3D shapes as voxels without actual 3D supervision, but requires many views for each object. Kato. *et al*. [11] proposed a differentiable renderer for meshes and reconstruct 3D meshes from images using the same principle. Tulsiani *et al*. [11] use a Differentiable Ray Consistency formulation in their learning framework to leverage different types of multi-view observations e.g. foreground masks, depth, color images, semantics etc. as supervision for learning single-view 3D prediction. Shape reconstruction from images is a challenging task due to the inherent ambiguity in a sparse set of views. Many approaches use task specific heuristics to design regularizers or constraints to deal with this problem. For example, the graph Laplacian of meshes was regularized [17, 9], and shapes were assumed to be symmetric [9]. Adversarial training is another way to learn shape priors. Wu *et al*. [19, 20] used discriminators on an estimated shape and its corresponding ground truth shape to make the estimated shapes more realistic. In contrast, our method does not require 3D models to learn prior knowledge. Our work is most similar to Kato *et al*. [10] and Gadelha *et al*. [5] both of which employ image based discriminators and do no use any 3D annotations.

3. Model

We aim to learn shape and novel view prediction systems, denoted as f_{shape} and f_{novel} respectively, which can

infer the corresponding property for the underlying object from a single image. However, instead of direct supervision, the superivison available is of the form of single or multi-view observations with their viewpoints annotations. We first formally define our problem setup by describing the representations inferred and training data leveraged and then discuss our approach.

Training Data. We require a sparse set of multi-view observations for multiple instances of the same object category. We assume a dataset of the form $\{\{(I_v^i, v_k^i) \mid k \in \{1, .., N_i\}\} \mid i \in \{1, .., N\}\}$. This corresponds to N object instances and N_i views available for the i^{th} instance. Associated with each image I_v^i, there is the viewpoint annotation v_k^i. Note that there is no explicit 3D supervision and the number of views available for an instance, denoted by N_i, can also be 1. In this paper, we focus on 3D shape learning by from silhouettes images, ignoring color, texture or lighting factors, thus representing each I_k^i as a binary two dimensional array.

Shape Representation. We use the volumetric representation of 3D shape V where each voxel V_i is a binary number. In other words, the voxel equals to one, i.e.$V_i = 1$, if the i^{th} voxel space is occupied by the shape; otherwise $V_i = 0$. The predicted shape V^p is a 3 dimensional occupancy grid, V^p, where the value of each grid cell (i, j, k) ranges between 0 and 1 representing the probability that the cell is occupied. To produce the final output we convert this occupancy grid into a binary grid by thresholding.

Projection. In this work, we assume orthographic projections and obtain the projected image from V^p with projection along the z-axis with the function $P((i, j), V) = 1 - \exp(\sum_k V(i, j, k))$. Intuitively, the operator sums up the voxel occupancy values along each line of sight (assuming othographic projection), and applies exponential falloff to create a smooth and differentiable function. When there is no voxel along the line of sight, the value is 0; as the number of voxels increases, the value approaches 1. To obtain a projection along a given viewpoint, v, we first rotate the volume by along the direction of the viewpoint, denoting the rotated volume as $V(v)$, and then apply the projection operator P. Combined with the rotated version of the voxel grid, we define our final projection module as: $P((i, j), V, v) = 1 - \exp(\sum_k V(v)(i, j, k))$.

3.1. Geometric consistency for shape

Multiple images of the same instance are simply renderings of a common geometry from different viewpoints. For the predicted and ground truth shapes to be similar their renderings, in our case the silhouettes, must also agree in the image space. Therefore, we minimize the difference between the projections of the predicted shape and ground truth views of the underlying object. Thus, the reconstruction loss is defined as:

$$L_{\text{data}} = \sum_{i=1}^{N} \sum_{j=1}^{N_i} \sum_{k=1}^{N_i} L(P(f_{\text{shape}}(I_j^i), v_k^i), I_k^i) \quad (1)$$

As our views are silhouette images we use the cosine distance as the loss function defined as:

$$L(I, \hat{I}) = 1 - \frac{I \cdot \hat{I}}{|I||\hat{I}|} \quad (2)$$

where $\cdot$ and $|.|$ represent the dot product and the L_2-norm operators respectively.

As described in section 1, with just one image per instance, the training objective is impervious to the appearance of the object from unobserved viewpoints. By imposing a geometrical constraint on the predicted shape with respect to a viewpoint not known a priori, we are compelling the reconstructor to build an implicit prior on shapes. Unless the network can memorize the set of auxiliary viewpoints for each input image, the predicted shapes will need to be reasonable from all possible perspectives for the loss to be small.

Consistency with novel views. We extend the applicability of the consistency principle to single view training by explicitly generating novel views from diverse viewpoints. Our approach acts as a data augmentation technique when multiple views are available in the data. In parallel to the shape reconstructor, f_{shape}, we use the *same* input image to predict the silhouette corresponding to a randomly sampled viewpoint. Then, we enforce that the predicted shape, when viewed according to this sampled viewpoint, should be consistent with the predicted novel view. Formally, using the shorthand V_j^i for the predicted shape for the i^{th} object from the j^{th} view, we write the total loss as:

$$L_{\text{data}} = \sum_{i=1}^{N} \sum_{j=1}^{N_i} \Bigg(\sum_{k=1}^{N_i} L\left(P(V_j^i, v_k^i), I_k^i\right) + L\left(P(V_j^i, \boldsymbol{w}), f_{\text{novel}}(I_j^i, \boldsymbol{w})\right) \Bigg) \quad (3)$$

$\boldsymbol{w}$ in the above equation is a random variable representing the novel viewpoint.

3.2. Novel view learning with GAN

We decouple the novel view generation objective into two parts: 1) there must be a common geometry that explains the input and the predicted views with their respective viewpoints, 2) the novel view should fit the distribution of ground truth renderings of the underlying object category.

Since their conception Generative Adversarial Networks [6] have become the dominant generative model due

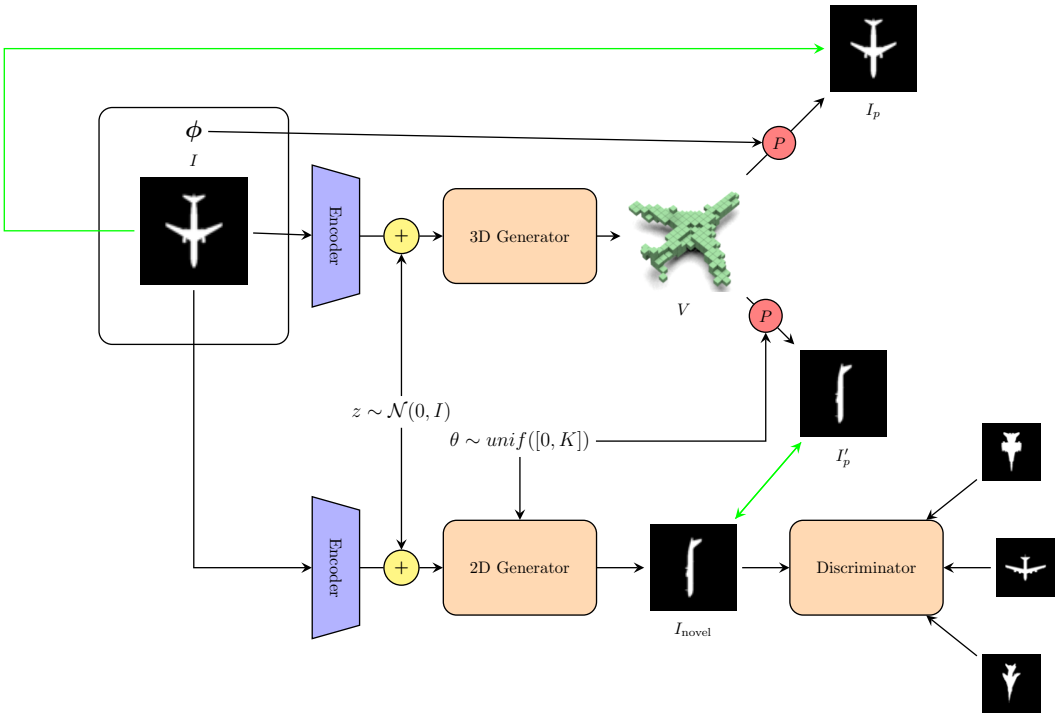

Figure 1: *The end-to-end architecture for training a 3D shape reconstructor with a single view per instance.* I is the input image and ϕ the corresponding viewpoint. V is a 3-dimensional occupancy grid and Π represents the projection operator. I_{novel} is the generated novel view conditioned on an encoding of the input image I and a randomly sampled viewpoint θ. I_p and I'_p are projections of V from viewpoints ϕ and θ respectively. The green arrows stand for direct image-to-image supervision such as using cosine distance between silhouettes. Notice the bidirectional arrows between I_{novel} and I'_p.

to its capcity to produce high quality images. Recent architectural [14, 15] and theoretical [1, 8, 13] investigations have contributed towards making GAN training more robust. As proposed in Goodfellow *et al.* [6], the Generative Adversarial Network (GAN) consists of a generator and a discriminator, where the discriminator tries to classify real objects and objects synthesized by the generator, and the generator attempts to confuse the discriminator. In a Generative Adversarial Network, the generator G maps a d-dimensional latent vector z to a $n \times n$ image, representing an image $G(z)$. The discriminator D outputs a confidence value $D(x)$ of whether a given input image $\tilde{I}$ is real or synthetic. To improve the quality and diversity of the results, we use the Wasserstein distance of WGAN-GP [1, 8]. Formally, we play the following minimax two-player game between G and D: $\min_G \max_D \mathcal{L}^{\text{GAN}}$, where

$$\mathcal{L}^{\text{GAN}} = \mathbb{E}_v[D(v)] - \mathbb{E}_z[D(G(z))]. \tag{4}$$

To enforce the Lipschitz constraint in Wasserstein GANs [1], we add a gradient penalty loss $L_{\text{GP}} = \lambda_{\text{GP}}\mathbb{E}_{\tilde{v}}[(\Delta_{\tilde{v}}D(\tilde{v}) - 1)^2]$ to equation 4, where $\tilde{v}$ is a randomly sampled point along the straight line between a real shape and a generated shape, and λ_{GP} controls the capacity of D. In our novel view GAN, the latent vector z is an encoding of the input image I produced by the encoder E_{novel} as represented in fig 1.

We further refine the space of produced images by utilizing the available viewpoint annotations in a conditinonal GAN architecture as proposed by Mirza and Osindero [12]. Let K represent the total number of discrete viewpoints in our data, we produce a viewpoint-conditional response from the generator by appending a one-hot K-dimensional vector to z. To condition the discriminator we append append the viewpoint vector along the channel dimension of real and generated images. Now we can define the generator and discriminator loss functions as follows:

$$\mathcal{L}_G^{\text{GAN}} = \sum_{i=1}^{N}\sum_{j=1}^{N_i}\left(D(G(E^{\text{novel}}(I_j^i)|\boldsymbol{w})\right). \tag{5}$$

$$\mathcal{L}_D^{\text{GAN}} = \sum_{i=1}^{N}\sum_{j=1}^{N_i}\left(D(I_j^i|v)) + L_{\text{GP}}\right). \tag{6}$$

Geometric consistency. We now formulate a loss function for geometric consistency between the predicted and the input image. Using the predicted shape as a proxy for the

underlying ground truth geometry, we define this loss as the distance, $L\left(f_{\text{novel}}(I_j^i, \boldsymbol{w}), P(V_j^i, \boldsymbol{w})\right)$.

This term is the same as the second term under summation in 3. Although the distance metric is commutative, we deliberately exchange the first and the second arguments to highlight a reversal in the roles of the predicted and the expected quantity. When observed from the shape prediction point of view, the (predicted) novel view acts as a supervisory signal; while in the present context, the predicted shape provides supervision for novel view generation. By adding this loss, the loss for the generator becomes:

$$\mathcal{L}_G^{\text{GAN}} = \sum_{i=1}^{N}\sum_{j=1}^{N_i}\Big(D(G(E^{\text{novel}}(I_j^i)|\boldsymbol{w})+ \lambda L\left(f_{\text{novel}}(I_j^i, \boldsymbol{w}), P(V_j^i, \boldsymbol{w})\right)\Big) \quad (7)$$

3.3. Implementation Details

Architecture. f_{shape} is a simple encoder decoder architecture. The encoder is a 5-layered of convolutional layers, each followed by ReLU and batch-norm, with an initial filter size of 64 doubling each layer. At the end, a fully connected layer maps the CNN output to an encoding of size 1024. We append a Gaussian noise vector of size 16 to this encoding to form the encoded input for the shape decoder. This noise captures the ambiguity around shape inherent in a single image. The shape decoder has 5 layers of deconvolutional. The first layer maps the $1040 \times 1 \times 1 \times 1$ size encoding to a $512 \times 4 \times 4 \times 4$ vector. Each subsequent deconv layer doubles the size along the three dimensions and halves the filters. All deconv layers except the last are followed by batch-norm and leaky ReLU. The last layer uses sigmoid to produce an occupancy grid with values between 0 and 1.

f_{novel} consists of a GAN, which has an encoder-decoder architecture, and another encoder to for the input image. The image encoder is a replica of the encoder from f_{shape}. The discriminator follows the same convolutional architecture, except the linear layer at the end produces a scalar output. The generator is similar in architecture to the shape decoder with 3d deconvolutional layers replaced by their 2d equivalent. The result is a 64×64 binary image.

Training. Using a mini-batch size of 64, our model takes 64×64 images as input and produces voxel grids with dimensions $64 \times 64 \times 64$. We train our model end-to-end by optimizing the three loss functions: shape loss 3, novel image generator loss 7 and discriminator loss 6. We apply one update to minimize each loss during every training iteration. Thus, we train the discriminator with the same frequency as the generator in our novel view GAN. We did it not find it useful to update the discriminator multiple times during each generator update. We tuned the value of λ in equation 7 lightly, and set it to 10 for all our experiments.

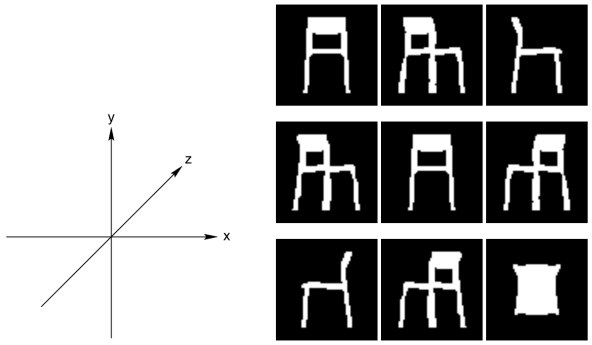

Figure 2: Illustration of viewpoints in our dataset for a chair. The first eight views are obtained by serially rotating around the y-axis by an angle of $45°$. For the last view, the object is rotated $90°$ along the x-axis.

4. Experiments

We use the ShapeNet dataset [2] to empirically validate our approach. We demonstrate that our model is able to produce reasonable 3d shapes and novel views through qualitative results for a number of object categories. Using multi-view consistency to train only f_{shape} as the baseline, we then present a comprehensive quantitative analysis.

Dataset. We evaluate on five commonly used categories with large number of models: plane, table, chair, car and firearm. Following the rules of orthographic projection, we render nine silhouette images for each model. The first eight of these are along the eight pre-selected directions evenly spaced around the y-axis (i.e. $0°, 45°, 90°, ..., 315°$), as seen in Fig 2. We obtain the last view from a $90°$ rotation of the volume along the x-axis.

4.1. Single-view training

We first tested our model in the single-view training set up. For each training instance, we randomly select one out of nine views and use this subset of the dataset for training. We compare our results with baseline that does not use novel view generation. In figure **??**, we show the results of our approach side-by-side with the baseline trained with single and 2 views per object instance. The baseline with single view uses the same data as our model for training; while, the two-view baseline requires an additional image per training instance.

In the single-view baseline, the network only learns to predict shapes that agree with the input image and no other prior on shape is enforced by the loss function. This explains the presence of artifacts in the produced voxels of all three categories. For example, there are superfluous voxels on the top and the bottom of the second *plane* which would not affect the projection from the input viewpoint. Similarly, the holes in the surface of the *chairs* lead to the same

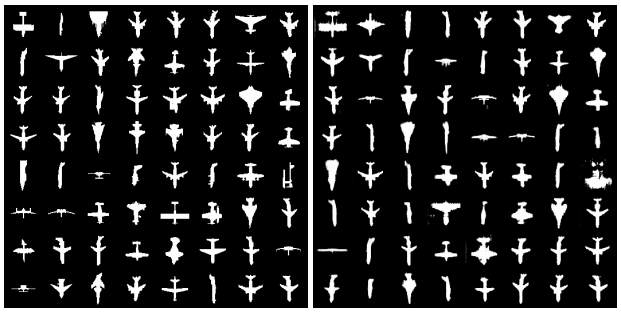

Figure 3: Novel view generation results; on left, the input ground truth images from a single batch; on right, the generated novel views.

projection as a regular surface. In contrast, our method resolves the ambiguity present in the silhouette images such that the produced shapes fit the desired distribution of these objects. This shows that even with a single image per object our model is able to build a reasonable prior on shapes.

Qualitatively, our results look comparable to those of the two-view baseline. As expected, multi-view training produces slightly better results as the supervision that we get from novel view generation in our model is replaced by ground truth supervision when another view is available. Still, our model provides a significant improvement over the single-view baseline in the data constrained setting.

4.2. Quantitative Results

For a quantitative analysis of our method we perform single view reconstruction and compute the mean intersection-over-union (IoU) as the evaluation metric. We experiment with single and multiple ground truth view supervision. We compare our results with the baseline which does not have a novel-view GAN. Table 1 lists results of these experiments. The addition of novel-view GAN provides a significant improvement over the baseline when only one view per object is available. Our model continues to help even with the addition of more ground truth views, but the improvement in performance over the baseline is not as dramatic. While in the single-view setting our model makes it possible to learn shape priors with the use of geometric consistency, with multiple views our approach of adding novel view supervision acts more like a data augmentation technique. Thus, with 2 or 3 ground truth views already available for learning a prior, the quality of the generated views becomes a crucial factor for improving improve 3D reconstruction. This explains the performance gap between our method with n ground truth views and the baseline with $n + 1$ views.

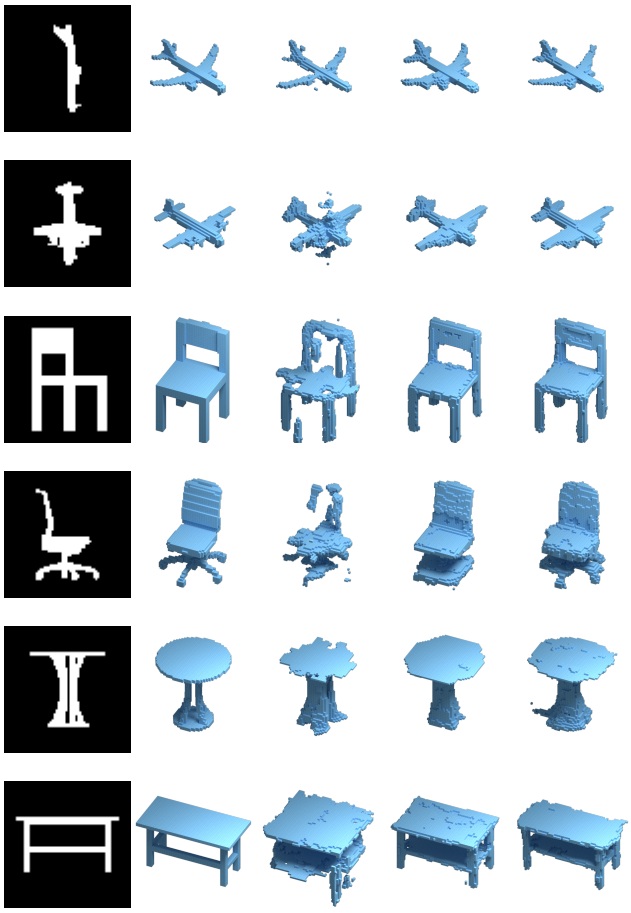

Figure 4: From left to right: Input image, ground truth 3D model, baseline using 1 view, ours using 1 view, baseline method with 2 views.

N_{gt}	*GAN*	*plane*	*chair*	*table*	*car*	*firearm*
1	✗	0.428	0.300	0.296	0.566	0.366
1	✓	0.496	0.379	0.377	0.677	0.456
2	✗	0.558	0.438	0.408	0.720	0.489
2	✓	0.566	0.455	0.441	0.729	0.479
3	✗	0.567	0.447	0.450	0.737	0.487
3	✓	0.573	0.451	0.460	0.742	0.488

Table 1: IoU comparison with baseline for different number of ground truth views.

5. Conclusion

In this work, we proposed a method to learn single image 3D reconstruction under a weakly supervised setting with images and their respective viewpoints. We verified our approach in single-view and multi-view settings for synthetic image datasets. The key to our success involves augmenting the 3D generator with a novel-view GAN in a way that

both of these components have complimentary objectives and the output of one component is used to provide an extra supervisory signal to the other. Our method significantly improves reconstruction accuracy over the baseline, especially in single-view training. This is a useful contribution because it is easier to create a single-view dataset than to create a multi-view dataset. This fact may enable 3D reconstruction of diverse objects beyond the existing synthetic datasets. A limitation of our approach is that we demonstrated its success assuming a dataset with precise viewpoint annotations. Moreover, we also assumed availability of silhouettes which can be tricky to obtain unless images with clean backgrounds are available. Our model, as presented in this work, is also incapable of reasoning about occlusions. These more general settings provide promising directions for further extension of our work. Although in this paper we demonstrate the potential of our method for the voxel representation, it is straightforward to extend the ideas to meshes and point clouds using an appropriate differentiable rendering engine.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. 4

[2] A. X. Chang, T. A. Funkhouser, L. J. 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. *CoRR*, abs/1512.03012, 2015. 5

[3] 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

[4] H. Fan, H. Su, and L. J. Guibas. A point set generation network for 3d object reconstruction from a single image. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2463–2471, 2017. 2

[5] M. Gadelha, S. Maji, and R. Wang. 3d shape induction from 2d views of multiple objects. In *2017 International Conference on 3D Vision (3DV)*, pages 402–411, Oct 2017. 1, 2

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. 3, 4

[7] T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry. A papier-mache approach to learning 3d surface generation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. 2

[8] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of wasserstein gans, 2017. 4

[9] A. Kanazawa, S. Tulsiani, A. A. Efros, and J. Malik. Learning category-specific mesh reconstruction from image collections. *Lecture Notes in Computer Science*, page 386402, 2018. 2

[10] H. Kato and T. Harada. Learning view priors for single-view 3d reconstruction, 2018. 1, 2

[11] H. Kato, Y. Ushiku, and T. Harada. Neural 3d mesh renderer. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. 1, 2

[12] M. Mirza and S. Osindero. Conditional generative adversarial nets. *ArXiv*, abs/1411.1784, 2014. 4

[13] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. 4

[14] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015. 4

[15] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *ArXiv*, abs/1606.03498, 2016. 4

[16] S. Tulsiani, T. Zhou, A. A. Efros, and J. Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. 1, 2

[17] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. *Lecture Notes in Computer Science*, page 5571, 2018. 2

[18] J. Wu, Y. Wang, T. Xue, X. Sun, W. T. Freeman, and J. B. Tenenbaum. MarrNet: 3D Shape Reconstruction via 2.5D Sketches. In *Advances In Neural Information Processing Systems*, 2017. 2

[19] 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 *Advances in Neural Information Processing Systems*, pages 82–90, 2016. 2

[20] J. Wu, C. Zhang, X. Zhang, Z. Zhang, W. T. Freeman, and J. B. Tenenbaum. Learning 3D Shape Priors for Shape Completion and Reconstruction. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[21] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision, 2016. 1, 2

[22] Y. Yang, C. Feng, Y. Shen, and D. Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. 2

[23] X. Zhang, Z. Zhang, C. Zhang, J. B. Tenenbaum, W. T. Freeman, and J. Wu. Learning to Reconstruct Shapes from Unseen Classes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2