Domain-Adaptive Single-View 3D Reconstruction

Pedro O. Pinheiro
Element AI

Negar Rostamzadeh
Element AI

Sungjin Ahn
Rutgers University

Abstract

Single-view 3D shape reconstruction is an important but challenging problem, mainly for two reasons. First, as shape annotation is very expensive to acquire, current methods rely on synthetic data, in which ground-truth 3D annotation is easy to obtain. However, this results in domain adaptation problem when applied to natural images. The second challenge is that there are multiple shapes that can explain a given 2D image. In this paper, we propose a framework to improve over these challenges using adversarial training. On one hand, we impose domain confusion between natural and synthetic image representations to reduce the distribution gap. On the other hand, we impose the reconstruction to be 'realistic' by forcing it to lie on a (learned) manifold of realistic object shapes. Our experiments show that these constraints improve performance by a large margin over baseline reconstruction models. We achieve results competitive with the state of the art with a much simpler architecture.

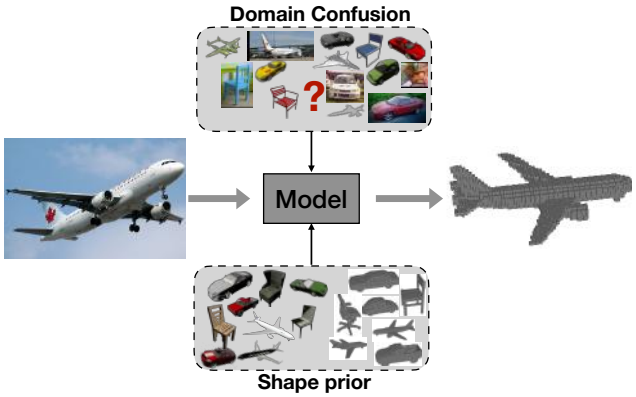


Figure 1: We propose a framework for (natural) single-view 3D reconstruction exploiting adversarial training in two ways. These constraints are achieved with additional loss terms. We impose domain confusion between natural and rendered images (top) and exploit shape priors to force reconstructions to look realistic (bottom).

1. Introduction

Humans can easily understand the underlying 3D structure of scenes and objects from single images. This is a hallmark of a human visual system and it is an essential step towards higher level visual understanding. This is an extremely ill-posed problem because a single image does not contain enough information to allow 3D reconstruction. Therefore, a machine vision system needs to rely on priors over the shape to infer 3D structure.

Efficient and effective 3D prototyping plays an important role in many different fields, such as virtual/augmented reality, architecture, robotics and 3D printing to name a few. Perhaps more importantly, studying 3D object representations could bring insights on how this information is encoded in intermediate and higher-level visual cortices [53, 26].

Traditional reconstruction methods rely on multiple images of same object instance [28, 4, 6, 39, 14]. These methods possess two strong limitations due to some key assumptions [8]: (i) it requires a large number of views to achieve reconstruction, (ii) the objects' appearance are expected to be Lambertian (*i.e.*, non-reflective) and their albedos are supposed to be non-uniform (*i.e.*, rich of non-homogeneous textures).

Another way to achieve 3D reconstruction is to leverage knowledge from object's appearance and shape. The main advantages of relying on *shape priors* is that we do not need to rely on accurate feature correspondences across different views. In this case 3D reconstruction can, in principle, be done from a single-view 2D image (assuming the priors are rich enough).

Recently, there has been a growing interest in learning-based approaches to tackle the problem of predicting the canonical shape of an object from a single image [24, 8, 16, 41, 54, 22, 48, 33, 44, 47, 49, 55]. Two technical advances were responsible for this surge: (i) the easy access to large-scale 3D Computer-Aided Design (CAD) repositories, such as ShapeNet [7], Pascal3D+ [52], ObjectNet3D [51], Pix3D [40] and (ii) advances in deep learning techniques [17].

Most of these methods contain a similar high-level archi-

tecture that regresses a 3D shape from (rendered) images: an encoder transforms a 2D image into a latent representation and a decoder reconstructs the 3D representation. They differentiate in how constraints from 3D world are imposed, *e.g.*, [8, 54, 44] force multi-view consistency to learn the 3D representation, while [47, 49] make use of 2.5D sketches. These approaches use a large number of CAD models to leverage shape priors (either making explicit use of 3D representation or not).

Single-view 3D reconstruction is a very ill-posed problem. In order to learn strong shape priors to infer 3D structure, deep learning methods require a large amount of 3D object annotations. However, acquiring good 3D object annotation from natural images is an extremely challenging endeavor. Most deep learning approaches, therefore, make use of synthetic images (which can be rendered easily if a proper 3D representation is given).

Convolutional neural networks (CNNs) [29] are known to perform sub-optimally when the data distribution of inputs changes, a problem known in the computer vision literature as *domain shift* [43]. For this reason, CNN-based 3D reconstruction, trained on synthetic images, performs worse when applied to natural images.

In this paper, we introduce a method to improve the performance of reconstruction models in natural images, where proper 3D labels are very difficult to acquire. To achieve this goal, we impose two constraints on the network's reconstruction loss (expressed as additional loss terms) based on shape prior learned from large 3D CAD repository (see Figure 1).

First, inspired by the domain adaptation literature [9, 15], we force the encoded 2D features to be invariant with respect to the domain they come from (rendered or natural). This way, a decoder trained on synthetic images will naturally perform better on real images. Second, we constraint the encoded 2D features to lie in the manifold of realistic objects shapes. This constraint forces the decoded 3D reconstruction to look more realistic. These two loss terms are characterized through adversarial training [18, 15], an active research topic.

Our main contributions can be summarized as follows: (i) we propose a model and a loss function that exploit learned shape priors to improve performance of natural image 3D reconstructions (using adversarial training in two different ways), (ii) we show that this method boost performance in both voxel and point cloud representations, and (iii) the proposed method achieves results competitive with state of the art on different datasets, with a much simpler architecture. Moreover, the proposed approach is independent of the encoder-decoder architecture and can be applied to different single-view 3D reconstruction models.

The rest of the paper is organized as follows: Section 2 presents related work, Section 3 describes how we learn the shape prior and leverage it in two different ways for learning reconstruction, and Section 4 describes our experiments in different datasets. We conclude in Section 5.

2. Related Work

Single-view 3D reconstruction. Traditional reconstruction methods rely on multiple images of same object instance to achieve reconstruction [28, 4, 6, 39, 14]. Recently, data-driven approaches to 3D reconstruction from single image have appeared. These methods can roughly be divided into two types: (i) those that explicitly use 3D structures [16, 8, 48, 13, 19, 47, 50] and (ii) those that use other sources of information to infer the 3D structure [46, 24, 54, 22, 20, 6, 44, 55].

These approaches, based on deep learning techniques, usually share a similar (high-level) architecture: an encoder that maps 2D (rendered) images into a latent representation and a decoder that maps this representation into a 3D object. They tend to differ in the way 3D world constraints are imposed. For instance, [8, 54, 54, 44, 20, 22, 27] force multiview consistency to learn the 3D representation, while [46, 24, 23] leverage keypoints and silhouette annotations. Other approaches [47, 49] leverage 2.5D sketches (surface normals, depth and silhouette) information to improve prediction.

More recently, Zhang, Zhang *et. al.* [56] consider spherical maps (in additional to 2.5D sketches) to learn 3D representations. Contrary to most work on single-view 3D reconstruction, the proposed method does not use canonical shape: every ground-truth 3D representation is on the same viewpoint as the 2D training sample. This work is the first to look at reconstructing shapes for unseen classes, however, it does not deal with domain-adaptation issues.

Contrary to all these methods, our approach does not use any additional information besides RGB images. However, in addition to rendered images, we also use unlabeled natural images (which are easy to acquire). We note that our contributions are independent of the encoder and decoder architecture (as long as they are differentiable), and could be applied in many of these more powerful encoder-decoder architectures. In experiments, we show that our approach improves performance over two baselines: a simple voxel encoder-decoder architecture and AtlasNet [19], a state-of-the-art encoder-decoder architecture based on point clouds representation.

Domain adaptation. The difficulty to acquire 3D annotations for natural images forces reconstruction models to learn from rendered images. It is well known in the literature [43, 9] that the performance of a model drops if applied in data coming from a distribution different from the one used during training. Ganin *et al.* [15] deal with this issue by forcing domain confusion (between two domains)

through an adversarial objective. Many works have been dealing with domain adaptation from synthetic to real for image classification [36, 37, 34, 38].

In this work, we borrow ideas from domain adaptation literature to impose domain confusion in a similar way as these previous work. We consider, however, the more challenging problem of 3D reconstruction instead of simple image classification.

Shape priors. Reconstruction of 3D structure from single-view images requires strong priors about object's shape. Many works focus on better capturing the manifold of realistic shapes. Non-deep approaches had focus on low-dimensional parametric models [3, 24].The authors of [16, 30] use CNNs to learn a common embedding space for 2D rendered images and 3D shapes. Other methods rely on generative modeling to learn shape prior, *e.g.*, [50] use deep belief nets to model 3D representations, [22, 6, 11] consider variants of variational autoencoders and [48] use a variant of GANs [18] to capture the manifold of shapes. In [31], the authors propose an adversarial autoencoder that uses adversarial training techniques to match aggregated posterior to perform variational inference.

A few works use adversarial training for single-view 3D reconstruction. Gwak *et al.* [20] use GANs to model 2D projections instead of 3D shapes. More similar to our work, Wu, Zhang *et al.* [49] use adversarial training techniques to impose reconstructions to look more natural. They use the discriminator of a pre-trained 3D GAN [48] to determine whether a shape is realistic. This approach is similar in principle to one of our contributions. It is, however, implemented in very different way. The input to the discriminator is a high dimensional 3D shape, which makes the training to be very unstable. In our method, the input is a single vector in a low-dimensional space.

3. Method

In our reconstruction setting, we are interested in predicting a *volumetric representation* $v^n \in \mathcal{V}$ from a canonical view of a natural image $x^n \in \mathcal{I}^n \subset \mathbb{R}^{3\times H\times W}$. In our experiments, the volumetric representation is either voxel ($\mathcal{V} \subset \{0,1\}^{d_v\times d_v\times d_v}$) or point cloud ($\mathcal{V} \subset \mathbb{R}^{d_v\times 3}$).

At training time, we have access to a large repository of 3D CAD objects, where pairs of rendered images and volumetric representation $\mathcal{D}_{rend} = \{(x_i^r, v_i)\}_{i=1}^{N_r}$ are drawn from a distribution $p_r(x, v)$, and unlabeled natural images, $\mathcal{D}_{nat} = \{x_j^n\}_{j=1}^{N_n}$, from a different distribution $p_n(x, v)$. We note that during training the model has access to natural images (which are easy to acquire), but not their voxel occupancy grid (which are very difficult to gather).

The proposed method, dubbed *Domain-Adaptive REConstruction network* (DAREC), is composed of two components: (i) a *shape autoencoder*, responsible for learning a rich latent representation of 3D objects and (ii) a *reconstruction network*, responsible for inferring the voxel occupancy grid from a 2D image.

The shape autoencoder is made of an encoder E and a decoder D. The encoder maps 3D representation $v \in \mathcal{V}$ into a low-dimensional embedding representation $e \in \mathcal{E} \subset \mathbb{R}^{d_e}$. The decoder maps a data point in the latent space back to a 3D representation. The voxel shape autoencoder is trained by minimizing the L_2 reconstruction loss. The point cloud shape autoencoder is trained by minimizing the Chamfer distance between predicted and ground truth points.

Since the shape autoencoder is trained with true 3D shapes, the learned latent representation lies in the *shape manifold* $\mathcal{E}$, containing low-dimensional embeddings of 'realistic' shapes. This component is trained prior to the training of the reconstruction network. The shape prior information is implicitly encoded in this rich representation space.

The reconstruction network also possesses an encoder-decoder architecture. The encoder f, parameterized by θ_f, is responsible to transform a 2D image into an embedding space from which a 3D representation can be reconstructed with a decoder. At inference time, the reconstruction network is the sole network used to predict the voxel occupancy of a given natural test image.

The model is trained in a way that the encoder mapping $f : \mathcal{I} \to \mathcal{E}$ can, at the same time: (i) reconstruct a 3D representation given a rendered image, (ii) be indistinguishable w.r.t. the domain that the image comes from (either synthetic or real) and (iii) stay in the manifold of 'realistic' shapes (learned with the shape autoencoder). To impose these constraints, we define and add the relevant terms to the loss function. Figure 2 shows an overview of the approach.

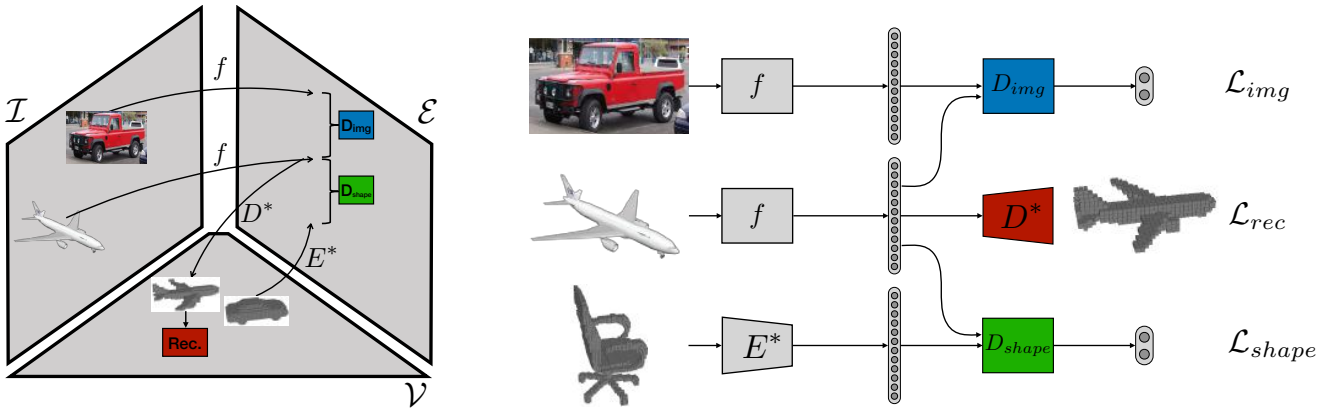
The reconstruction loss, $\mathcal{L}_{rec}$, is applied to tuples of rendered images and 3D representations (from $\mathcal{D}_{rend}$). We use the L_2 for reconstruction loss when considering voxel representation and the Chamfer distance (as in [13, 19]) for the point cloud representation. We opt to *not* update the decoder parameters at this training stage. This design choice, combined with the constraint imposed by the third loss, forces the image representations to lie on the manifold of 'realistic' shapes.

In the rest of this section, we show how we leverage adversarial training techniques and (learned) shape prior to improve performance of natural image 3D reconstruction.

3.1. Confusing Image Domains

It is well known that machine learning algorithms suffer from domain shift [43]. Therefore, a model trained to reconstruct 3D shape from rendered images performs sub-optimally when applied to natural ones.

Theoretical studies [2, 1] suggest that a good cross-domain representation is one in which input domain can-



(a) (b)

Figure 2: (a) The reconstruction network maps an image to a rich embedding space $\mathcal{E}$ which is then decoded into a 3D shape with the shape decoder D^* (the star indicates we do not update the parameters of D and E). The two constraints are imposed on the embedding space with the help of two discriminators: D_{img} imposes image domain confusion and D_{shape} forces the embeddings to lie on the shape manifold. (b) Overview of the proposed architecture.

not be easily identified. We implement such domain confusion by mapping cross-domain features into a common space through adversarial training. We cast this problem as a minimax game between a domain classifier and feature encoders. That is, we encourage the feature encoder to learn features $f(x)$ that maximize the domain confusion between natural and rendered images.

We consider a discriminator D_{img}, parameterized by θ_{img}. The discriminator that classifies the domain of a given feature vector, is optimized by the standard adversarial classification loss as follows:

$$\mathcal{L}_{img}(\theta_f, \theta_{img}) = -\mathop{\mathbb{E}}_{x^r \sim p_r} \log D_{img}(f(x^r)) + \\ -\mathop{\mathbb{E}}_{x^n \sim p_n} \log\left(1 - D_{img}(f(x^n))\right). \quad (1)$$

We achieve domain confusion by applying Reverse Gradient algorithm [15], which optimizes the parameters θ_f to maximize the discriminator loss directly, while θ_{img} minimizes it.

3.2. Exploiting Shape Priors

A lot of inherent ambiguity exists in single image reconstruction. Multiple objects exists that can explain a single view. For this reason, as noted by Wu, Zhang *et al.* [49], 3D reconstruction with only supervised loss tends to predict unrealistic mean shapes.

We characterized a representation to be 'realistic' if it belongs close to the manifold created by the (learned) shape autoencoder. We argue that if the feature of a single 2D image $f(x)$ lies in the same manifold, a realistic reconstruction can be achieved by leveraging the decoder of the shape autoencoder.

The third component of our loss, $\mathcal{L}_{shape}$, imposes this constraint by penalizing the model if the distribution of latent embeddings does not match that of the points in the shape manifold. We rely on the learned shape autoencoder to sample these points. Again, we use adversarial training to optimize the loss.

Similar to Equation 1, we train a discriminator D_{shape} (parameterized by θ_{shape}) to classify whether a sample is drawn from a 2D encoding representation or from the shape manifold. Samples from the shape manifold are generated by sampling voxel (or point cloud) instances from ShapeNet and mapping them to the $\mathcal{E}$, using the learned shape encoder E^*. The star means that the parameters of the encoder are kept unchanged during this stage of training. This way, we guarantee the encoded samples lie on the learned manifold.

Learning is achieved by minimizing the following loss:

$$\mathcal{L}_{shape}(\theta_f, \theta_{shape}) = -\mathop{\mathbb{E}}_{x^r \sim p_r} \log D_{shape}(f(x^r)) + \\ -\mathop{\mathbb{E}}_{v \sim p_r} \log\left(1 - D_{shape}(E^*(v^r))\right). \quad (2)$$

As before, the parameters θ_{shape} are optimized to minimize this loss while the parameters θ_f maximize it, therefore, forcing the 2D embeddings to lie on the shape manifold.

3.3. Training Details

The training procedure is done in two stages.

We start by training the shape autoencoder to learn shape priors. As we want to capture the intrinsic shape complexity of different objects, we train the model using the full

ShapeNet dataset. We use a different shape autoencoder for each 3D representation considered.

The voxel autoencoder has an encoder E composed of four 3D convolutional layers, each followed by a max-pooling and ReLU [32] non-linearity. The first layer contains 5×5 filters while the remaining have 3×3. The number of hidden units are 32, 64, 128 and 256 respectively. Similarly, the voxel decoder D has four convolution layers, but instead of max-pooling, we use bilinear upsampling. The dimension of the latent representation is 256.

We use AtlasNet [19] [1] for the point cloud autoencoder. The encoder, similar to PointNet [35], transforms the input point cloud into a latent representation of dimension 1024. The decoder contains four fully-connected layers of size 1024, 512, 256, 128 with ReLU non-linearities (except the last layer, which has a tanh).

Once training converges, we freeze the parameters of the encoder and the decoder and use them in the reconstruction step.

The architecture of the reconstruction network is shown on Figure 2b. The parameters of network f are initialized with a ResNet-50 [21] that was pre-trained to perform classification on ImageNet dataset [10]. We replace the classification layer by a randomly initialized layer that outputs a vector with dimension of the latent space.

The two discriminators D_{img} and D_{shape} map the embedded features to the probability of which domain the input comes from (modeled by a softmax [5]). We use two fully-connected layers of dimension 1024, followed by ReLU. We choose not to share but have different set of parameters between the two discriminators because it performs sightly better in practice.

Finally the model is optimized to learn 3D representations that are domain-invariant and that lie in the manifold from the prior of realistic shapes. Consequently, our final goal is to optimize the following objective:

$$\begin{aligned} \min_{\theta_f} \max_{\theta_{img}, \theta_{shape}} \quad & \mathcal{L}_{rec}(\theta_f) \\ & - \lambda_i \mathcal{L}_{img}(\theta_f, \theta_{img}) \\ & - \lambda_s \mathcal{L}_{shape}(\theta_f, \theta_{shape}) \ , \end{aligned} \tag{3}$$

where λ_i and λ_s are balance parameters between the loss terms. We chose λ_i and λ_s to be both 0.001 when considering voxel representation and 0.01 with point cloud representation. To optimize, we used Adam [25] with learning rate of 10^{-4} for voxel and 10^{-5} for point cloud representation.

4. Experiments

In this section, we start by comparing the performance of our approach with other methods on the problem of single-view reconstruction from natural images. We report results with two variants of the model: DAREC-vox, which predicts voxel representations and DAREC-pc, which predicts point cloud representations. We evaluate the models in two important datasets: the recently released Pix3D [40] and PASCAL 3D+ [52]. Then, we study how DAREC behaves with respect to the different loss terms. Finally, we analyze the learned representation and show qualitative results that corroborates with the notion of domain confusion and shape manifold.

4.1. Experimental Setup

Voxels and point clouds. In the first stage of training, we learn shape priors by training a shape autoencoder for the two 3D representations considered. In both cases, the autoencoder is trained to reconstruct the shape (voxels or point clouds) from ShapeNet dataset [7] (we use the ShapeNetCore subset). This dataset contains over 50k object instances of 55 categories. We use a voxel resolution of 32^3 (a downsampled version of the voxels provided by the official repository) and 2500 points in the point cloud representation.

The second training stage is responsible for inferring shape representation from a single-view image. We train two versions of our model: (i) DAREC-vox, which outputs voxel representations and uses the voxel autoencoder and (ii) DAREC-pc, which regresses point clouds and uses AtlasNet for the point cloud shape autoencoder. In this step, we make use of both natural and rendered 2D images. We follow previous work and use the same rendered view provided by [8]. This allow a more fair comparison between the proposed method and other approaches. Since we evaluate the model in natural images, we use all rendered data for training.

Evaluation metrics. We evaluate the performance of our method using two metrics: Intersection over Union (IoU) and Chamfer Distance (CD). The metric IoU measures the similarity between ground-truth and (discretized) reconstruction voxels. This is the 3D extension of the common metric (of same name) used in segmentation. The Chamfer distance between two point clouds $P_1, P_2 \subset \mathbb{R}^3$ is defined as:

$$CD(P_1, P_2) = \frac{1}{|P_1|} \sum_{x \in P_1} \min_{y \in P_2} ||x - y||_2 + \frac{1}{|P_2|} \sum_{x \in P_2} \min_{y \in P_1} ||x - y||_2. \tag{4}$$

For each point in each set, CD finds the closest point (in the other set) and average the distances. When dealing with voxel occupancy, we first sample points in the the voxel isosurface before computing CD. It is shown by Sun, Wu *et al.* [40] that CD better correlates with human perception. For fair comparison, in the following sections we use the same evaluation code provided by the authors of Pix3D[2].

[1] we use the official code provided at https://github.com/ThibaultGROUEIX/AtlasNet

[2] http://pix3d.csail.mit.edu/

	IoU	CD
3D-R2N2[8]	0.136	0.239
3D-VAE-GAN [48]	0.171	0.182
PSGN* [13]	-	0.199
MarrNet† [47]	0.231	0.144
DRC† [44]	0.265	0.160
AtlasNet [19]	-	0.148
AtlasNet + g.t. mask* [19]	-	0.126
ShapeHD† [49]	0.284	0.123
DAREC-vox	0.241	0.140
DAREC-pc	-	0.112

Table 1: Single-view 3D reconstruction results on Pix3D. We show results on both IoU and CD metrics. * PSGN require ground-truth mask as input. † MarrNet, DRC and ShapeHD use 2.5D sketches to guide training. Our approach only considers (easily available) natural images during training. We show competitive results in both metrics.

4.2. Comparison to Other Methods

Reconstruction on Pix3D. Pix3D is a large-scale benchmark of diverse image-shape pairs with pixel-level 2D-3D alignment. A significant part of the dataset is chairs because they are common and highly diverse. Following the previous works [40, 49], we evaluate our approach on the 2,894 untruncated and unoccluded 'chair' images.

During training, the reconstruction network has access to synthetic ShapeNet renderings (and their corresponding ground-truth, voxels or point clouds) and unlabeled natural images of 'chair' category (we use the natural images of the PASCAL 3D+ r1.1, which contains also ImageNet images). Figure 3 shows the qualitative results of voxel reconstructions generated by our approach. As illustrated in this figure, DAREC is able to reconstruct even in situations of strong self-occlusion.

Table 1 compares the performance of our approach with different methods on the Pix3D dataset. We show results on both IoU (higher is better) and CD (lower is better) metrics. Results from other models are taken from Wu, Zhang *et al.* [49].

It is also important to mention that these methods use different types of data during training. For instance, PSGN [13] require ground-truth masks as input. MarrNet [47], DRC [44] and ShapeHD [49] use depth, surface normals and silhouettes during training. DAREC achieves competitive results using only RGB images as input and with a much simpler architecture.

Reconstruction on Pascal 3D+. PASCAL 3D+ [52] provides annotations for (rough) 3D shape of different rigid object instances from PASCAL VOC 2012 [12]. Each category has a small set of about 10 CADs per category.

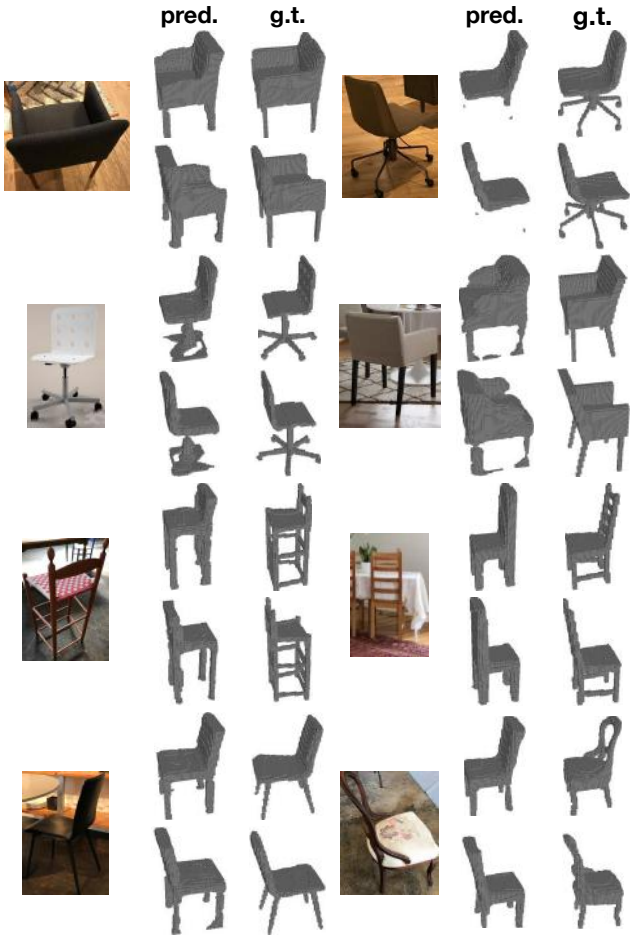


Figure 3: 3D reconstruction from single image on Pix3D dataset. For each image, we show the predicted and the ground-truth voxel representations. Our method is capable of learning shape with very different appearances. We show two different views for each 3D representation.

Similar to most of the recent works, we do not use any of the PASCAL 3D+ training set. We use the CAD annotations only for benchmarking purposes. As discussed in Tulsiani *et al.* [44], using the small set of CADs for both training and test would bias the model toward those samples and therefore is not a recommended benchmark protocol.

We train our model in the categories that are present in both Pascal3D+ and ShapeNet renderings provided by [8]: 'aeroplane','car','chair','table' and 'tv monitor'. During training, our approach uses ShapeNet rendered images-shape tuples and natural images (we use natural images from ImageNet [10]). Figure 4 shows voxel reconstruction results on different images (and their corresponding ground-truths).

Table 2 shows the performance (in terms of CD) of different methods. Following previous work [44, 49], we show

	chair	car	plane	average
3D-R2N2[8]	0.238	0.305	0.305	0.284
DRC [44]	0.158	0.099	0.112	0.122
OGN [42]	-	0.087	-	-
ShapeHD [49]	0.137	0.129	0.094	0.119
DAREC-vox	0.135	0.101	0.108	0.115
DAREC-pc	0.140	0.100	0.112	0.117

Table 2: Single-view 3D reconstruction results on Pascal3D+. We show results on CD metrics. DRC and ShapeHD use depth/normals/silhouettes as extra information during training. OGN considers a much stronger decoder and much higher voxel resolution. Our approach only considers (easily available) natural images.

results in three categories. Our approach achieves comparable state-of-the-art results with both 3D representations. As before, our reconstruction network, contrary to other methods, does not make use of depths, surface normals, silhouette nor it exploits any form of multiview consistency. Instead, we make use of unlabeled natural images, which are very easy to obtain. We also note that OGN [42] uses a much more complex decoder (based on octrees) and considers much higher resolution volumetric occupancy ground-truths during training.

We note that the two novelties of our approach are complementary to previous works and thus could potentially be integrated with those methods for further performance gain.

4.3. Analyzing the Loss

Here, we perform an ablation study to see how our method performs with respect to different loss terms. Table 3 shows results of our method on Pix3D chair dataset when considering: (i) only the reconstruction loss ($\mathcal{L}_{rec}$), (ii) the reconstruction and the shape prior losses ($\mathcal{L}_{rec}$ and $\mathcal{L}_{shape}$), (iii) the reconstruction and the image domain-confusion losses ($\mathcal{L}_{rec}$ and $\mathcal{L}_{img}$) and (iv) the full loss. In all those cases, the models have same capacity at inference time (all of them consists of same encoder/decoder architecture).

The first row ($\mathcal{L}_{rec}$ only) ignores the adversarial losses and is trained only with synthetic images. Our method is able to improve the performance by a large margin with both 3D representations (*e.g.*, from .220 to .140 with voxel and .148 to .112 with point cloud).

We first observe that, for both datasets, each loss term has a positive impact on the the final reconstruction result. The shape prior loss alone is not sufficient to significantly improve the performance. However, the domain confusion loss alone already provides a substantial boost in performance. Finally, the model achieves its best performance when combining both constraints at the same time.

			Pix3D	
$\mathcal{L}_{rec}$	$\mathcal{L}_{img}$	$\mathcal{L}_{shape}$	voxel	point cloud
✓			.220	.148
✓		✓	.196	.140
✓	✓		.156	.129
✓	✓	✓	.140	.112

Table 3: The performance of our model, considering different loss terms, measured with CD on Pix3D chair datasets. We note the importance of each loss component on both metrics, although the shape prior loss alone does not give considerable improvement.

These results therefore confirm that each of the proposed loss terms is critical in obtaining the final performance.

4.4. Analyzing the Learned Representations

Feature visualization. We use t-SNE [45] to visualize feature representations from different domains and at different adaptation stages (we use the DAREC-vox model on Pix3D). Figure 5(a-b) shows t-SNE features from synthetic (blue) and real (red) images before and after adaptation, respectively. Figure 5(c-d) shows embeddings (before and after training, respectively) of 2D rendered images (blue) and points from the learned shape manifold, *i.e.*, latent representations from the shape autoencoder (yellow).

In both cases, we can see that features become much more domain-invariant after training, as desired. During our experiments, we indeed observed a strong correspondence between reconstruction performance (on natural images) and the overlap between the different feature distributions.

Shape interpolation. In Figure 6, we show results of interpolating between two natural images of different shapes. We first transform each image into its latent representation. Then, we walk through the shape manifold and reconstruct the shape at different interpolated representations. We show qualitatively that the learned shape manifold gives smooth transition between the two object shapes.

Shape arithmetic. Another way to probe the learned representations is to show arithmetic on the latent space. Previous work [8, 48, 55] showed they are able to learn a semantic manifold of shapes in its latent space and arithmetic is done in samples from this space. In Figure 7, we perform shape arithmetic on different natural images. We first map them to the learned shape manifold (where arithmetical operations are done), then we reconstruct its shape. We observe that the representation after the arithmetic operations are still reasonable to reconstruct a realistic shape.

Figure 4: 3D reconstruction from single image on PASCAL 3D+. For each image, we show the predicted and ground-truth (left) voxel representation (right). We show two different views for each 3D representation.

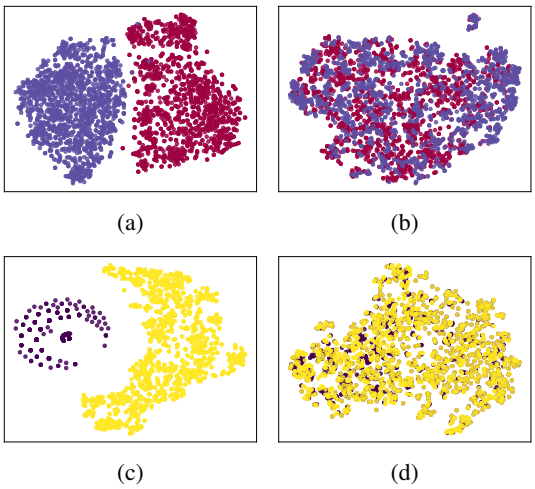


Figure 5: t-SNE visualization on Pix3D. (a-b) Rendered and natural image embeddings before and after domain confusion. (c-d) 2d rendered embedding and points from manifold before and after training.

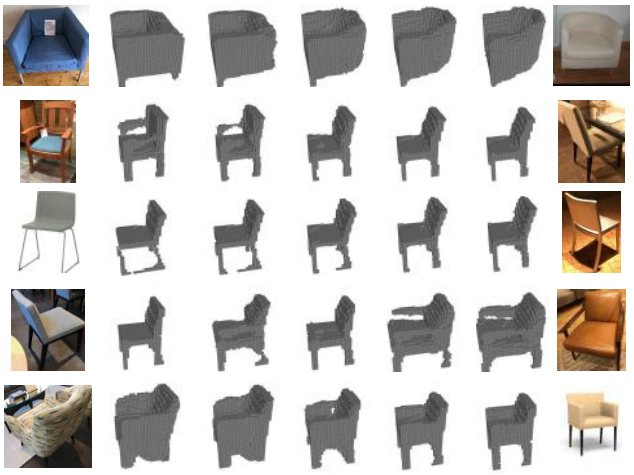


Figure 6: Shape interpolation from natural images.

Figure 7: Shape arithmetic from natural images. The top row show that 'curviness' vector can be added to other chairs. The other rows show that 'arm' vector can be added to other chairs.

5. Conclusion

In this paper, we presented a framework for improved 3D reconstruction from single-view natural image. Our method leverages adversarial training and shape priors in two different ways. First it imposes learned features to be domain-invariant to help with the problem of domain adaptation. Second, we force the learned representations to lie in a rich shape prior manifold, imposing the reconstructions to be realistic. We show our method is able to improve the performance when considering different 3D representations. By using only RGB signal and with a much simpler network architecture, our model achieves competitive performance with the state of the art.

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 2010. 3

[2] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *NIPS*, 2007. 3

[3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 1999. 3

[4] Jeremy S. De Bonet and Paul A. Viola. Roxels: Responsibility weighted 3d volume reconstruction. In *ICCV*, 1999. 1, 2

[5] John Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. *Neurocomputing: Algorithms, Architectures and Applications*, 1990. 5

[6] Adrian Broadhurst, Tom Drummond, and Roberto Cipolla. A probabilistic framework for space carving. In *ICCV*, 2001. 1, 2, 3

[7] Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. In *CoRR*, 2015. 1, 5

[8] Christopher Bongsoo Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *ECCV*, 2016. 1, 2, 5, 6, 7

[9] Gabriela Csurka. A comprehensive survey on domain adaptation for visual applications. In *Domain Adaptation in Computer Vision Applications*, Advances in Computer Vision and Pattern Recognition. Springer, 2017. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5, 6

[11] S. M. Ali Eslami, Danilo Jimenez Rezende, Frédéric Besse, Fabio Viola, Ari S. Morcos, Marta Garnelo, Avraham Ruderman, Andrei A. Rusu, Ivo Danihelka, Karol Gregor, David P. Reichert, Lars Buesing, Theophane Weber, Oriol Vinyals, Dan Rosenbaum, Neil C. Rabinowitz, Helen King, Chloe Hillier, Matthew M Botvinick, Daan Wierstra, Koray Kavukcuoglu, and Demis Hassabis. Neural scene representation and rendering. *Science*, 2018. 3

[12] Mark Everingham, S. M. Eslami, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 2015. 6

[13] Haoqiang Fan, Hao Su, and Leonidas J. Guibas. A point set generation network for 3d object reconstruction from a single image. In *CVPR*, 2017. 2, 3, 6

[14] Yasutaka Furukawa and Carlos Hernández. Multi-view stereo: A tutorial. *Foundations and Trends in Computer Graphics and Vision*, 2015. 1, 2

[15] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016. 2, 4

[16] Rohit Girdhar, David F. Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *ECCV*, 2016. 1, 2, 3

[17] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. The MIT Press, 2016. 1

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 2, 3

[19] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan C. Russell, and Mathieu Aubry. Atlasnet: A papier-mâché approach to learning 3d surface generation. 2018. 2, 3, 5, 6

[20] JunYoung Gwak, Christopher B. Choy, Animesh Garg, Manmohan Chandraker, and Silvio Savarese. Weakly supervised generative adversarial networks for 3d reconstruction. In *3DV*, 2017. 2, 3

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[22] Danilo Jimenez Rezende, S. M. Ali Eslami, Shakir Mohamed, Peter Battaglia, Max Jaderberg, and Nicolas Heess. Unsupervised learning of 3d structure from images. In *NIPS*, 2016. 1, 2, 3

[23] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 2

[24] Abhishek Kar, Shubham Tulsiani, João Carreira, and Jitendra Malik. Category-specific object reconstruction from a single image. In *CVPR*, 2015. 1, 2, 3

[25] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2014. 5

[26] Zoe Kourtzi and Charles E Connor. Neural representations for object perception: structure, category, and adaptive coding. *Annual review of neuroscience*, 2011. 1

[27] Abhijit Kundu, Yin Li, and James M. Rehg. 3d-rcnn: Instance-level 3d object reconstruction via render-and-compare. In *CVPR*, 2018. 2

[28] A. Laurentini. The visual hull concept for silhouette-based image understanding. *PAMI*, 1994. 1, 2

[29] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, 1998. 2

[30] Yangyan Li, Hao Su, Charles Ruizhongtai Qi, Noa Fish, Daniel Cohen-Or, and Leonidas J. Guibas. Joint embeddings of shapes and images via cnn image purification. *ACM Trans. Graph.*, 2015. 3

[31] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian Goodfellow. Adversarial autoencoders. In *ICLR*, 2016. 3

[32] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010. 5

[33] David Novotný, Diane Larlus, and Andrea Vedaldi. Learning 3d object categories by looking around them. In *ICCV*, 2017. 1

[34] Pedro O. Pinheiro. Unsupervised domain adaptation with similarity learning. In *CVPR*, 2018. 3

[35] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017. 5

[36] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Adversarial dropout regularization. In *ICLR*, 2018. 3

[37] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 2018. 3

[38] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *ECCV*, 2018. 3

[39] Steven M. Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR*, 2006. 1, 2

[40] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B. Tenenbaum, and William T. Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *CVPR*, 2018. 1, 5, 6

[41] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Multi-view 3d models from single images with a convolutional network. In *ECCV*, 2016. 1

[42] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *ICCV*, 2017. 7

[43] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011. 2, 3

[44] Shubham Tulsiani, Tinghui Zhou, Alexei A. Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *CVPR*, 2017. 1, 2, 6, 7

[45] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *JMLR*, 2008. 7

[46] Sara Vicente, João Carreira, Lourdes Agapito, and Jorge Batista. Reconstructing pascal voc. In *CVPR*, 2014. 2

[47] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum. Marrnet: 3d shape reconstruction via 2.5d sketches. In *NIPS*, 2017. 1, 2, 6

[48] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *NIPS*, 2016. 1, 2, 3, 6, 7

[49] Jiajun Wu, Chengkai Zhang, Xiuming Zhang, Zhoutong Zhang, William T. Freeman, and Joshua B. Tenenbaum. Learning shape priors for single-view 3d completion and reconstruction. In *ECCV*, 2018. 1, 2, 3, 4, 6, 7

[50] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, 2015. 2, 3

[51] Yu Xiang, Wonhui Kim, Wei Chen, Jingwei Ji, Christopher Choy, Hao Su, Roozbeh Mottaghi, Leonidas Guibas, and Silvio Savarese. Objectnet3d: A large scale database for 3d object recognition. In *ECCV*, 2016. 1

[52] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *WACV*, 2014. 1, 5, 6

[53] Yukako Yamane, Eric T Carlson, Katherine C Bowman, Zhihong Wang, and Charles E Connor. A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nature Neuroscience*, 2008. 1

[54] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *NIPS*, 2016. 1, 2

[55] Guandao Yang, Yin Cui, Serge Belongie, and Bharath Hariharan. Learning single-view 3d reconstruction with limited pose supervision. In *ECCV*, 2018. 1, 2, 7

[56] Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Josh Tenenbaum, Bill Freeman, and Jiajun Wu. Learning to reconstruct shapes from unseen classes. In *NeurIPS*. 2018. 2