

# Single-view Object Shape Reconstruction Using Deep Shape Prior and Silhouette

Kejie Li<sup>12</sup>

kejie.li@adelaide.edu.au

Ravi Garg<sup>12</sup>

ravi.garg@adelaide.edu.au

Ming Cai<sup>12</sup>

ming.cai@adelaide.edu.au

Ian Reid<sup>12</sup>

ian.reid@adelaide.edu.au

<sup>1</sup> The School of Computer Science  
University of Adelaide  
Adelaide, Australia

<sup>2</sup> Australian Centre for Robotic Vision  
Adelaide, Australia

---

## Abstract

3D shape reconstruction from a single image is a highly ill-posed problem. Modern deep learning based systems try to solve this problem by learning an end-to-end mapping from image to shape via a deep network. In this paper, we aim to solve this problem via an online optimization framework inspired by traditional methods. Our framework employs a deep autoencoder to learn a set of latent codes of 3D object shapes, which are fitted by a probabilistic shape prior using Gaussian Mixture Model (GMM). At inference, the shape and pose are jointly optimized guided by both image cues and deep shape prior without relying on an initialization from any trained deep nets. Surprisingly, our method achieves comparable performance to state-of-the-art methods even without training an end-to-end network, which shows a promising step in this direction.

## 1 Introduction

Humans effortlessly infer the 3D structure of objects from a single image, including the parts of the object which are not directly visible in the image. It is obvious that the human perception of object geometry involves, in large part, the vast experience of having seen the objects of a category from multiple-views, having reasoned about their geometry and having built rich prior knowledge about meaningful object structure for a specific object category. However, single view object shape reconstruction remains a challenging problem for modern vision algorithms – perhaps because of the difficulty of capturing and representing this prior shape information – and the solutions to 3D geometric perception have been traditionally dominated by multi-view reconstruction pipelines.

The recent successes of deep learning have motivated researchers to depart from traditional multi-view reconstruction methods and directly learn the image to object-shape mapping using CNNs [10, 18, 21, 42, 45, 49, 50]. These deep networks show impressive results, with prior information about shape captured at training time and represented implicitly in the network’s weights. However, the lack of real images with corresponding ground-truth

3D shapes forces these methods mostly to rely on synthetic images rendered from a object CAD model repository for training, which obviously leads to a domain gap when evaluating on real images.

An alternative way to reason about the 3D geometry of the object is by combining the 2D image cues (i.e. object silhouette) with explicitly learned priors about the object shape – thereby separating the 2D recognition from 3D reconstruction. This two-stage framework has been successfully used for inferring 3D shape from images prior to the resurgence of deep learning by [57, 58] given a single view and [9, 14] in a multi-view scenario. They propose to estimate the 3D object shape and pose given silhouette(s). To make this highly ill-posed problem well-posed, these works learned a probabilistic latent space embedding using a Gaussian Process Latent Variable Model (GPLVM). An image cue based energy function is minimized with respect to the object shape and pose.

These methods worked quite well for object categories with small shape variations and which lack high frequency detail – like cars or human bodies. Nevertheless, more complex objects like chairs, with larger shape variation and thin structures, were significant failure cases. Some of the limitations arise due to the use of GPLVM for dimensionality reduction. GPLVM does not scale well to be trained on large datasets or to learn priors on high-dimensional shape representations. This forces [58] to use frequency based compression techniques for the representing shapes leading to loss of thin and other high frequency object structures.

Throughout this paper, we aim to answer one important question in the field of single-view object shape reconstruction: Is learning an end-to-end deep network using enormous number of synthetic images really necessary? Although they show promising results in recent years, not only they suffer from domain gap, but also they are expensive to train in terms of synthetic image rendering and network training time. More importantly, one common drawback of deep neural network is the non-transparency property being a black box, which leads to a problem that geometry constraints (i.e. silhouette consistency and photometric consistency) that traditional approaches use for geometry reconstruction may not be enforced at inference regardless of whether those constraints are used during training.

To answer the question aforementioned, we present an approach that takes advantage of both deep learning and traditional optimization based methods. Technically, we train a point cloud deep autoencoder that captures a latent space of object shapes, which is further modeled by a Gaussian Mixture Model (GMM). At inference, an object shape is optimized in the form of latent code by jointly maximizing the probability explained by the GMM and minimizing the discrepancy between its projection to image plane using object pose and an object silhouette estimated from an off-the-shelf semantic segmentation network (so-called silhouette consistency). Note that unlike single-view 3D reconstruction, a semantic segmentation network can easily be trained on real-image dataset (e.g. MSCOCO [54]), thereby domain gap is avoided.

Although some approaches [50, 54, 55] also tried to enforce silhouette consistency at inference time, our method is significantly different from them in twofold. Firstly, we initialize the shape optimization using the mean latent codes of the GMM, whereas previous approaches rely on a pre-trained image-to-shape deep network for initialization. Secondly, additional to the silhouette consistency, the search of object shape is constrained by a shape prior term via the GMM, which empirically shows its efficacy in our ablation study. The proposed approach not only achieves comparable performance with the state-of-the-art end-to-end network, but also enjoys the advantage of simplicity and flexibility. An illustration of our framework can be found on Figure 1.

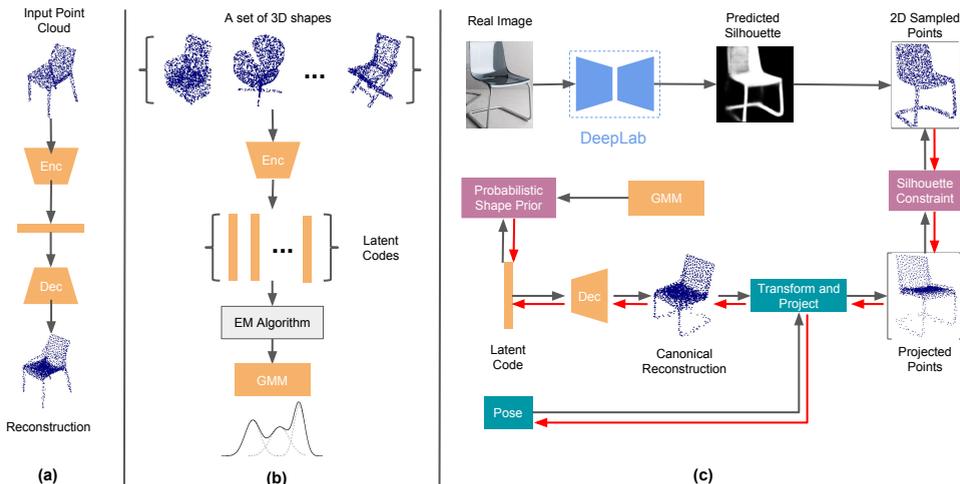


Figure 1: (a) We train the point cloud autoencoder to learn a latent space for 3D shapes (b) A set of 3D point clouds are mapped to the latent codes by the trained point cloud encoder. A GMM as a probabilistic model of the latent space is learned by fitting on this set of latent codes. (c) At inference time, we jointly estimate the shape and pose guided by the probabilistic shape prior via the learned GMM and the silhouette-shape constraint. The **orange**, **green**, **blue**, and **purple** blocks correspond to the components for shape, pose, silhouette, and loss terms respectively. The red arrows  $\rightarrow$  represent the gradient flow with respect to the latent code and the object pose during inference.

In summary, our contributions are as follows:

- By leveraging the power of modern deep learning for non-linear dimensionality reduction, we show that a simple GMM fitting into this latent space can effectively be a probabilistic shape prior that plays a crucial role in our optimization framework.
- While pioneering works [50, 55] show that online optimization can further improve the one-shot network prediction, to the best of our knowledge, we present the first method that does not rely on a deep network to initialize the shape and pose optimization, which makes our method simpler.
- Even without training an end-to-end network, our shape-prior aware inference framework achieve comparable performance to deep learning based prior art for single-view object shape reconstruction on a real image benchmark dataset.

## 2 Related Works

3D understanding of object shapes from images is considered an important step in scene perception. While mapping an environment using Structure from Motion [3, 43] and SLAM [15, 36] facilitates localization and navigation, a higher level of understanding about objects in terms of their shape and relative position with respect to the rest of the background is essential for their manipulation. Initial works, to localize objects while estimating its pose

from an image, are limited to the case where a pre-scanned object structure is available [19, 40, 41].

Soon it is realized that the valid 3D shapes of objects belonging to a specific category are highly correlated and dimensionality reduction emerges as a prominent tool to model object shapes. Works like [46] reconstruct traditional image segmentation datasets like PASCAL VOC [17] by extending the ideas of non-rigid Structure from Motion [6, 12, 13, 20]. Methods like [29] propose to use these learned category specific object shape manifolds to reconstruct shape of the object from single image by fitting the reconstructed shape to a single image silhouette and refine it with simple image based methods like shape from shading [24].

While the methods mentioned above attempt to learn linear object shape manifolds from 2D key-point correspondences or object segmentation annotations, increasing availability of class specific 3D CAD models and object scans allow researchers to learn complex manifolds for shapes within an object category directly from data. For example, Gaussian Process Latent Variable Model or Kernel/simple Principal Component Analysis is employed to learn a compact latent space of object shapes in [13, 14, 16, 58]. Taking advantage of modern deep learning techniques, deep auto-encoders [25, 47], VAEs [50], and GANs [22, 49] can be better alternatives to model object shapes than traditional dimensionality reduction tools.

Recently, with the advance of deep learning, there are many end-to-end systems proposed to solve the object shape reconstruction problem from a single RGB image directly. Different supervision has been explored. Direct 3D supervision is dominant in the early works [11, 18, 21, 23, 50]. Weaker supervision that uses geometry projection (i.e. depth consistency, silhouette consistency) is later employed in [22, 27, 32, 53, 44, 45, 48, 53, 54] as one step moving toward training the network using real images. However, all multiple views of an object are required to train these networks, and they still rely on a one-shot prediction from the network at inference time.

Most related to our work are CodeSLAM [6], MarrNet [50], and Zhu *et al.* [53], which also use a low-dimensional latent space learnt by a deep network to represent the geometry information. For the scene-level geometry reconstruction, CodeSLAM uses nearby views to form a photometric loss, which is minimized by searching in the learned latent space of depth maps. MarrNet [50] and Zhu *et al.* [53] enforce the 2.5D (silhouette, depth or normal) - 3D shape constraint and photometric loss respectively to search in the latent space of object shapes. Our work differs from theirs in twofold. Firstly, a common disadvantage of these works, which we address in our work, is that they do not consider a probabilistic prior in the latent space explicitly when they perform optimization, which we show is essential for the optimization at inference time through our ablation study. Secondly, we do not need to train a deep network to initialize optimization variables (i.e. shape latent code and pose)

### 3 Method

The goal of this work is to reconstruct a 3D shape of an object from a single image. Our framework closely follows traditional object prior based 3D reconstruction methods by framing the reconstruction as an optimization problem assuming that a deep neural network provides us with the object silhouette. This inference method is described in the Section 3.1. To facilitate the inference, we require a low-dimensional latent space which encodes the object shape, a learned probabilistic prior on this latent space that estimates the confidence of a latent code. How we learn the relevant prerequisites for inference is elaborated in Section 3.2, followed by implementation details in Section 3.3.

### 3.1 Inference-as-optimization

For inference, we are given with sampled points from the predicted silhouette  $\mathcal{S}_{sil}$ , initial latent code of the object shape  $\mathbf{l}_0$ , initial pose  $\mathbf{R}_0$ . Additionally, the learned probabilistic prior on the latent space is available in the form of a GMM.

We minimize the objective function defined below with respect to the shape and pose variables  $\mathbf{l}$  and  $\mathbf{R}$ :

$$L(\mathbf{l}, \mathbf{R}) = L_{sil} + \lambda L_{shape}. \quad (1)$$

Here  $L_{sil}$  corresponds to the silhouette-shape constraint and  $L_{shape}$  is the term related to the probabilistic shape prior, which we explain in the rest of this section.  $\lambda$  is the weighting factor between these two loss terms.

**Silhouette-shape Constraint Term.** It enforces the projection of a transformed 3D shape to be consistent with the object’s silhouette on the image. While using point cloud representation, this can be achieved by defining a 2D Chamfer loss between projection of the transformed 3D point cloud and sampled points of the predicted silhouette shown in Equation (2),

$$\begin{aligned} \mathcal{S} &= dec(\mathbf{l}), \\ L_{sil}(\mathbf{P}, \mathcal{S}, \mathcal{S}_{sil}) &= \sum_{\mathbf{s}_i \in \mathcal{S}} \min_{\mathbf{s}_j \in \mathcal{S}_{sil}} \|\mathbf{P}\mathbf{s}_i - \mathbf{s}_j\|^2 + \sum_{\mathbf{s}_j \in \mathcal{S}_{sil}} \min_{\mathbf{s}_i \in \mathcal{S}} \|\mathbf{P}\mathbf{s}_i - \mathbf{s}_j\|^2, \end{aligned} \quad (2)$$

where  $\mathcal{S}$  is a point cloud reconstruction in a predefined canonical pose generated by a latent code  $\mathbf{l}$  through the point cloud decoder.  $\mathbf{P}$  is a transformation matrix that is composed of  $\mathbf{K}[\mathbf{R}]$ , projection matrix  $\mathbf{K}$ , object rotation  $\mathbf{R}$  (we use orthogonal projection because camera intrinsic is assumed unknown).  $\mathcal{S}_{sil}$  is a set of 2D sampled points from a predicted silhouette.

**Shape Prior Term.** Because of the highly unconstrained nature of the silhouette-shape constraint and incorrect silhouette, we employ a shape prior term, which quantifies the likelihood of a latent code using a probabilistic model. This can mitigate the problem of generating unrealistic 3D shapes. Specifically, the shape prior minimizes the negative log likelihood of a latent code in the GMM, which can be written as:

$$L_{shape}(\mathbf{l}) = -\log\left(\sum_{k=0}^K \pi_k \mathcal{N}(\mathbf{l} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right), \quad (3)$$

where  $\pi_k$ ,  $\boldsymbol{\mu}_k$ , and  $\boldsymbol{\Sigma}_k$  are the weight, mean, and covariance matrix of the  $k^{th}$  Gaussian component, and  $K$  is the number of Gaussian components.

**Initialization** A sensible initialization of the shape latent code is using one of the means of GMM. However, we observed that the optimized shape is susceptible to local minimums if we initialize the shape latent code from one position only due to the non-convexity of the optimization. We therefore run the optimization multiple times with different means of the GMM to alleviate this problem. We initialize the rotation at 0 degree for azimuth, elevation, and tilt.

### 3.2 Learning

Our inference method relies on learning a low-dimensional latent space of shape from which a decoder network can map a latent code  $\mathbf{l}$  to a unique 3D point cloud  $\mathcal{S}$ . To learn this mapping, we train an autoencoder. Similar to a conventional autoencoder, our encoder  $enc(\cdot)$

maps a point cloud  $\mathcal{S}$  to its corresponding low dimensional latent code  $\mathbf{l}$ , followed by the decoder  $dec(\cdot)$  reconstructs the point cloud  $\hat{\mathcal{S}}$  from the latent code  $\mathbf{l}$ . We train this autoencoder using 3D Chamfer Distance as defined below:

$$L_{cd}(\hat{\mathcal{S}}, \mathcal{S}) = \sum_{\hat{\mathbf{s}}_i \in \hat{\mathcal{S}}} \min_{\mathbf{s}_j \in \mathcal{S}} \|\hat{\mathbf{s}}_i - \mathbf{s}_j\|^2 + \sum_{\mathbf{s}_i \in \mathcal{S}} \min_{\hat{\mathbf{s}}_j \in \hat{\mathcal{S}}} \|\hat{\mathbf{s}}_j - \mathbf{s}_i\|^2. \quad (4)$$

To learn a probabilistic model as the shape prior, we fit a GMM to the learned latent space. Specifically, we map the point clouds  $\{\mathcal{S}_1 \dots \mathcal{S}_N\}$  to the latent space using the point cloud encoder, to get a set of latent codes  $\{\mathbf{l}_1 \dots \mathbf{l}_N\}$ . The parameters of GMM with  $K$  components are then learned to maximize the likelihood of the latent codes by using the EM algorithm to optimize the following objective:

$$\min_{\substack{\pi_1 \dots \pi_K, \\ \boldsymbol{\mu}_1 \dots \boldsymbol{\mu}_K, \\ \boldsymbol{\Sigma}_1 \dots \boldsymbol{\Sigma}_K}} - \sum_{i=0}^N \log \left( \sum_{k=0}^K \pi_k \mathcal{N}(\mathbf{l}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right), \quad (5)$$

where  $\boldsymbol{\pi}_k$ ,  $\boldsymbol{\mu}_k$ , and  $\boldsymbol{\Sigma}_k$  are the weight, mean and covariance matrix of the  $k^{th}$  Gaussian component, and  $N$  and  $K$  are the number of sample latent codes and the number of Gaussian component.

There are a number of generative models that we considered besides GMM during development. For instance, we experimented with VAE and GAN. However, we found that VAE fails to reconstruct accurately, which we suspect is due to two reasons. Firstly, one well-known issue of VAE is that the KL divergence is too restrictive [14], which makes the reconstruction uninformative to the input. Secondly, a simple isotropic gaussian prior may not be expressive for complicated and diverse shapes. Although the latent space learnt by GAN can produce realistic shapes, a problem we observe is mode collapse in local region of the latent space, which means that neighboring latent codes correspond to identical shape. It makes traversing in the latent space difficult for our gradient based optimization. Other drawbacks of using GAN are also revealed by [1], which empirically show that, in the form of point cloud, among variants of GAN, and GMM, the latter outperforms other models in generalization and diversity of shapes, meaning that the GMM captures a better distribution of object shapes.

### 3.3 Implementation Details

The architecture of our encoder is similar to PointNet [39]. We use five 1D convolution layers, followed by a max pooling as a ‘‘permutation invariance’’ function, and two fully connected layers, to map from a point cloud (a  $2048 \times 3$  matrix) to a latent code. The decoder is composed of five fully connected layers to map from a latent code to a point cloud reconstruction. ReLU [35] is used between layers except the last one as a non-linear activation function. Point clouds used for training are samples from the CAD models of the ShapeNet dataset [8]. The autoencoder is trained on 13 object classes including chair, table and sofa, and is trained for 150 epochs using Adam optimizer [40]. We use GMM with 5 Gaussian components for chair class and 4 Gaussian components for table and sofa class when evaluating on Pix3D [12]. The number of Gaussian components is chosen such that each mean represents a distinctive object shape. The GMM is optimized using the EM algorithm implemented in Python Scipy library [28].

Methods	CD ↓	EMD ↓
3D-R2N2	0.239	0.211
PSGN	0.200	0.216
3D-VAE-GAN	0.182	0.176
DRC	0.160	0.144
MarrNet	0.144	0.136
AtlasNet	0.125	0.128
<b>Ours</b>	<b>0.116</b>	0.125
Pix3D	0.119	<b>0.118</b>

Table 1: 3D shape reconstruction results on Pix3D dataset

At inference, the object silhouette is given by DeepLab v3 [10]. Rotation parameters and shape latent code are optimized jointly reusing Adam optimizer in Tensorflow [11].  $\lambda$  during the first 500 iterations is set to a relatively high value at 1 to reduce the shape-pose ambiguity, and later reduced to 0.001. The optimization is terminated when the rate of change in loss function diminishes or the maximum number of 1000 iterations is reached.

## 4 Experiments

We provide experimental results of our framework on single view object shape reconstruction on a real dataset and compare it against prior art. A series of ablation studies are also presented.

### 4.1 Testing Data and Evaluation Protocol

We evaluate on a recently published Pix3D [12] dataset – a dataset consisting diverse shapes aligned with corresponding real images. We benchmark with prior art on chair class following Pix3D setup, where 2894 real images and 221 corresponding 3D ground-truth shapes are used. We also showcase that our method can adapt to other object classes by experimenting on table and sofa class in the ablation study section. Other alternatives to Pix3D which were used popularly were either purely synthetic datasets like ShapeNet [8], or PASCAL 3D+ [52] which was created by associating a small number of overlapping 3D CAD models per category for both training and test images of a particular object category to create pseudo ground truth. Due to this reason, as discussed in [12, 50] this dataset is not deemed fit for shape reconstruction evaluation recently.

We use 3D Chamfer Distance (CD as defined in (4)) and Earth Mover Distance (EMD as defined in (6)) as evaluation measures for shape reconstruction following the human study on the evaluation metrics for shape reconstruction presented in Pix3D which shows that these matrices are more correlated with human perception than IoU.

$$EMD(S_1, S_2) = \frac{1}{\|S_1\|} \min_{\phi: S_1 \rightarrow S_2} \sum_{x \in S_1} \|x - \phi(x)\|_2, \quad (6)$$

where  $\phi: S_1 \rightarrow S_2$  is a bijection. We use the approximation implemented in Pix3D benchmark setup to calculate the EMD efficiently.

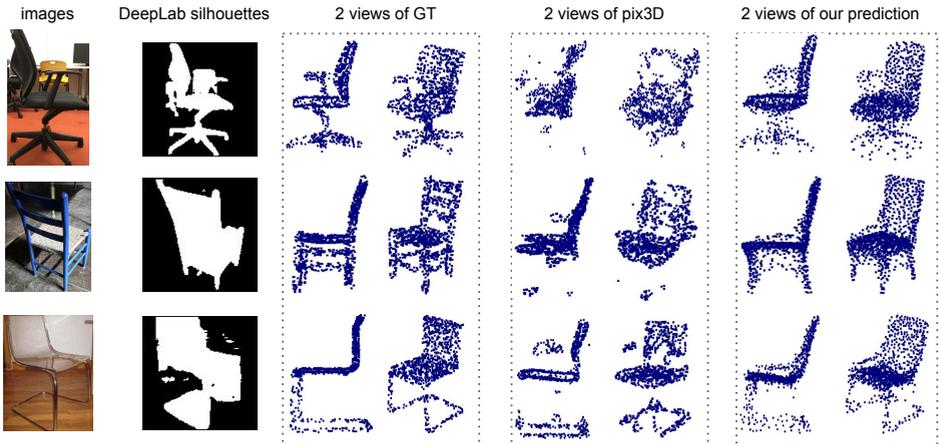


Figure 2: Qualitative comparison with Pix3D.

Loss	$L_{sil}$ only		$L_{sil} + L_{shape}$	
	CD	EMD	CD	EMD
chair	0.137	0.148	<b>0.116</b>	<b>0.125</b>
table	0.173	0.176	<b>0.158</b>	<b>0.172</b>
sofa	0.116	0.108	<b>0.099</b>	<b>0.092</b>

Table 2: Quantitative analysis on the effect of silhouette-shape constraint and probabilistic shape prior at inference.

## 4.2 Shape Reconstruction on Pix3D

We compare the performance of our 3D shape prediction pipeline with a list of prior art including 3D-VAE-GAN [49], 3D-R2N2 [10], PSGN [18], DRC [44], AtlasNet [23], MarrNet [60] and Pix3D [61] which is a variant of MarrNet that decouples shape and pose prediction. The quantitative result is reported in Table 1. The performance evaluation for most of these methods listed above is taken from Pix3D paper [47]. Despite the simplicity of our method, which only relies on deep shape prior and silhouette consistency within an optimization framework, we outperform many end-to-end deep networks that are specifically designed to reconstruct object shape from a single view. Qualitative results are shown in Figure 2 for visual comparison. We hope this result can facilitate the community to re-think the role of deep learning in the task of single-view object shape reconstruction or even other geometry reconstruction tasks. Is an end-to-end network really necessary or the best practice for single-view shape reconstruction?

## 4.3 Ablation Study on Shape prior

In this section, we demonstrate that the proposed probabilistic shape prior plays a crucial part in our optimization pipeline. To that end, we compare the optimization using silhouette only and our full loss term which takes into account the likelihood of generated shape by using both  $L_{sil}$  and  $L_{shape}$ .

As for the quantitative comparison in Table 2, when we minimize the objective function

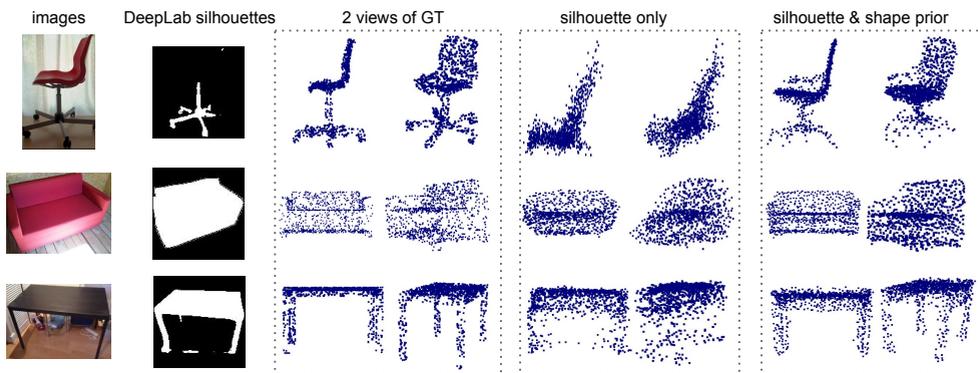


Figure 3: Qualitative results using GMM shape prior.

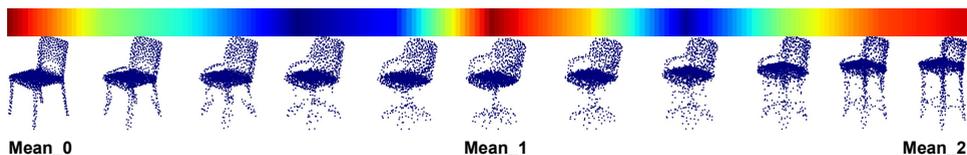


Figure 4: This figure visualizes that a “chair-like” shape has higher confident in the GMM. We denote the likelihood of a shape explained by the GMM using colors. Red means high likelihood whereas blue represents low likelihood.

that penalizes a latent code with low likelihood, the reconstruction performance improves on both measures significantly.

The main takeaway from these results is that, jointly optimizing shape and pose using silhouette only is a highly ill-posed problem and leads to unrealistic shape with a wrong pose in order to fit the silhouette. The GMM shape prior avoid the regions of the latent space to restrict the inferred shape to be limited to realistic object-like solutions – thereby avoiding the overfitting to a noisy silhouettes. We present Figure 3 for visual comparison.

To further explore this hypothesis, we choose three points from the latent space corresponding to the mean of different Gaussian components and interpolate linearly in the latent space to generate intermediate sample chair shapes. The results are visualized in Figure 4. It can be clearly seen that the latent space learned using deep autoencoder training constitutes of regions corresponding to very not chair-like shapes. However, these not chair-like shapes usually correspond to the high variance region of the learned GMM – for instance, between the latent code of a four-legged chair (mean\_0) and that of a swivel chair (mean\_1), the likelihood of an unrealistic shape as a mix of two structures is particularly low.

## 5 Discussion

Instead of training an end-to-end network to address single-view object shape reconstruction, we present an online optimization framework that optimizes the object shape in the form of deep latent code at inference without relying on any deep learning initialization. The latent search space is constrained by a shape prior via a GMM fitted in the latent space and object silhouette given by an off-the-shelf segmentation network, which can be trained with real

images and does not suffer from domain gap. This approach achieves comparable state-of-the-art performance on a large dataset of real images when enjoying its simplicity. While this result shows an interesting direction to tackle single-view object shape reconstruction besides training an end-to-end network, there are a number of directions that are worth investigation. For example, instead of fitting a GMM after a latent space is learnt, how to simultaneously learn a latent space and fit a multi-cluster probabilistic shape prior? It would also be interesting to apply our shape-prior aware optimization framework to a multi-view setup for reconstruction and tracking.

## References

- [1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL <https://www.tensorflow.org/>. Software available from tensorflow.org.
- [2] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas J Guibas. Learning representations and generative models for 3d point clouds. *arXiv preprint arXiv:1707.02392*, 2017.
- [3] Sameer Agarwal, Noah Snavely, Ian Simon, Steven M Seitz, and Richard Szeliski. Building rome in a day. In *International Conference on Computer Vision*, pages 72–79. IEEE, 2009.
- [4] S. Y. Bao, M. Chandraker, Y. Lin, and S. Savarese. Dense object reconstruction with semantic priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1264–1271, June 2013. doi: 10.1109/CVPR.2013.167.
- [5] Michael Bloesch, Jan Czarnowski, Ronald Clark, Stefan Leutenegger, and Andrew J Davison. Codeslam-learning a compact, optimisable representation for dense visual slam. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [6] W Brand. Morphable 3d models from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages II–II. IEEE, 2001.
- [7] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3d shape from image streams. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 690–696. IEEE, 2000.
- [8] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

- [9] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Re-thinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.
- [10] Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016.
- [11] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, pages 628–644. Springer, 2016.
- [12] Yuchao Dai, Hongdong Li, and Mingyi He. A simple prior-free method for non-rigid structure-from-motion factorization. *International Journal of Computer Vision*, 107(2): 101–122, 2014.
- [13] Samuel Dambreville, Yogesh Rathi, and Allen Tannenbaum. A framework for image segmentation using shape models and kernel space shape priors. *IEEE transactions on pattern analysis and machine intelligence*, 30(8):1385–1399, 2008.
- [14] Amaury Dame, Victor A Prisacariu, Carl Y Ren, and Ian Reid. Dense reconstruction using 3d object shape priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1288–1295, 2013.
- [15] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):1052–1067, 2007.
- [16] Francis Engelmann, Jörg Stückler, and Bastian Leibe. Joint object pose estimation and shape reconstruction in urban street scenes using 3d shape priors. In *German Conference on Pattern Recognition*, pages 219–230. Springer, 2016.
- [17] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.
- [18] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, page 6, 2017.
- [19] Juergen Gall, Bodo Rosenhahn, and Hans-Peter Seidel. Clustered stochastic optimization for object recognition and pose estimation. In *Joint Pattern Recognition Symposium*, pages 32–41. Springer, 2007.
- [20] Ravi Garg, Anastasios Roussos, and Lourdes Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1272–1279, 2013.
- [21] Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *European Conference on Computer Vision*, pages 484–499. Springer, 2016.

- [22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [23] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. Atlasnet: A papier-m<sup>^</sup>ach<sup>\</sup>e approach to learning 3d surface generation. *arXiv preprint arXiv:1802.05384*, 2018.
- [24] Paul Henderson and Vittorio Ferrari. Learning to generate and reconstruct 3d meshes with only 2d supervision. *arXiv preprint arXiv:1807.09259*, 2018.
- [25] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.
- [26] Berthold KP Horn. *Obtaining shape from shading information*. MIT press, 1989.
- [27] Eldar Insafutdinov and Alexey Dosovitskiy. Unsupervised learning of shape and pose with differentiable point clouds. In *Advances in Neural Information Processing Systems*, 2018.
- [28] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. URL <http://www.scipy.org/>. [Online; accessed <today>].
- [29] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. Category-specific object reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1966–1974, 2015.
- [30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [31] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [32] Kejie Li, Trung Pham, Huangying Zhan, and Ian Reid. Efficient dense point cloud object reconstruction using deformation vector fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 497–513, 2018.
- [33] Chen-Hsuan Lin, Chen Kong, and Simon Lucey. Learning efficient point cloud generation for dense 3d object reconstruction. In *AAAI Conference on Artificial Intelligence*, 2018.
- [34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [35] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.
- [36] Richard A Newcombe, Steven J Lovegrove, and Andrew J Davison. Dtam: Dense tracking and mapping in real-time. In *Proceedings of IEEE International Conference on Computer Vision*, pages 2320–2327. IEEE, 2011.

- [37] Victor Adrian Prisacariu and Ian Reid. Shared shape spaces. In *Proceedings of the 2011 International Conference on Computer Vision*, pages 2587–2594. IEEE Computer Society, 2011.
- [38] Victor Adrian Prisacariu, Aleksandr V Segal, and Ian Reid. Simultaneous monocular 2d segmentation, 3d pose recovery and 3d reconstruction. In *Asian conference on computer vision*, pages 593–606. Springer, 2012.
- [39] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1(2):4, 2017.
- [40] Bodo Rosenhahn, Thomas Brox, and Joachim Weickert. Three-dimensional shape knowledge for joint image segmentation and pose tracking. *International Journal of Computer Vision*, 73(3):243–262, 2007.
- [41] Christian Schmaltz, Bodo Rosenhahn, Thomas Brox, Joachim Weickert, Daniel Cremers, Lennart Wietzke, and Gerald Sommer. Occlusion modeling by tracking multiple objects. In *Joint Pattern Recognition Symposium*, pages 173–183. Springer, 2007.
- [42] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2974–2983, 2018.
- [43] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2): 137–154, 1992.
- [44] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 3, 2017.
- [45] Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Multi-view consistency as supervisory signal for learning shape and pose prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [46] Sara Vicente, Joao Carreira, Lourdes Agapito, and Jorge Batista. Reconstructing pascal voc. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 41–48, 2014.
- [47] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008.
- [48] Olivia Wiles and Andrew Zisserman. Silnet: Single-and multi-view reconstruction by learning from silhouettes. In *British Machine Vision Conference*, 2017.
- [49] 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 *Advances in Neural Information Processing Systems*, pages 82–90, 2016.

- [50] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum. Marrnet: 3d shape reconstruction via 2.5 d sketches. In *Advances in neural information processing systems*, pages 540–550, 2017.
- [51] Jiajun Wu, Tianfan Xue, Joseph J Lim, Yuandong Tian, Joshua B Tenenbaum, Antonio Torralba, and William T Freeman. 3d interpreter networks for viewer-centered wireframe modeling. *International Journal of Computer Vision*, pages 1–18, 2018.
- [52] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 75–82. IEEE, 2014.
- [53] Xinchun Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *Advances in Neural Information Processing Systems*, pages 1696–1704, 2016.
- [54] Rui Zhu, Hamed Kiani Galoogahi, Chaoyang Wang, and Simon Lucey. Rethinking reprojection: Closing the loop for pose-aware shape reconstruction from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 57–65. IEEE, 2017.
- [55] Rui Zhu, Chaoyang Wang, Chen-Hsuan Lin, Ziyang Wang, and Simon Lucey. Object-centric photometric bundle adjustment with deep shape prior. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 894–902. IEEE, 2018.