Deep Optimized Priors for 3D Shape Modeling and Reconstruction

Mingyue Yang[1]*, Yuxin Wen[1]*, Weikai Chen[2], Yongwei Chen[1], Kui Jia[134]†
[1]South China University of Technology, [2]Tencent Game AI Research Center
[3]Pazhou Laboratory, [4]Peng Cheng Laboratory
{eemingyueyang,wen.yuxin}@mail.scut.edu.cn,
chenwk891@gmail.com, eecyw@mail.scut.edu.cn, kuijia@scut.edu.cn

Abstract

Many learning-based approaches have difficulty scaling to unseen data, as the generality of its learned prior is limited to the scale and variations of the training samples. This holds particularly true with 3D learning tasks, given the sparsity of 3D datasets available. We introduce a new learning framework for 3D modeling and reconstruction that greatly improves the generalization ability of a deep generator. Our approach strives to connect the good ends of both learning-based and optimization-based methods. In particular, unlike the common practice that fixes the pre-trained priors at test time, we propose to further optimize the learned prior and latent code according to the input physical measurements after the training. We show that the proposed strategy effectively breaks the barriers constrained by the pre-trained priors and could lead to high-quality adaptation to unseen data. We realize our framework using the implicit surface representation and validate the efficacy of our approach in a variety of challenging tasks that take highly sparse or collapsed observations as input. Experimental results show that our approach compares favorably with the state-of-the-art methods in terms of both generality and accuracy.

1. Introduction

Deep generative models have brought impressive advances to the state-of-the-art across a wide variety of generative tasks, including 2D image synthesis and 3D shape reconstruction. At the moment, it is widely believed that these leaps in performance come primarily from the realistic priors learned from a large amount of training data. Based on this observation, most of the previous 3D learning approaches focus on learning stronger priors during training and strictly respecting the learned prior at test time. Specifically, there are two common ways to leverage the learned shape priors. One is to train an encoder to retrieve the most likely prior by mapping the input into the latent code, which is a low-dimensional representation of the shape prior. The other is to optimize the latent code until its decoded output achieves a minimal loss. Note that both methods fix the learned prior/generator once the training is completed as the prior is considered to be the most valuable asset in a learning-based approach. However, is this the best strategy of using the prior in a 3D learning task?

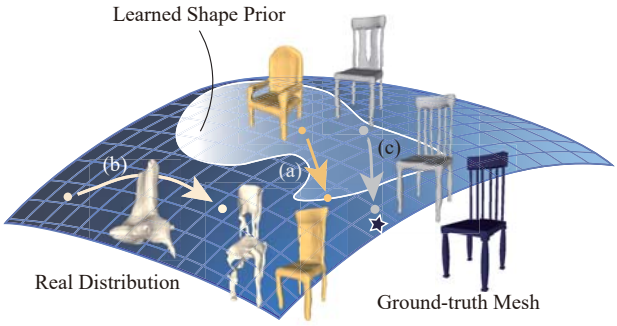

Figure 1: The shape prior learned from the limited training data cannot capture the full landscape of the real data distribution. Common practice that uses a fixed pre-trained generator is constrained within the prior (path a) and thus fails to model the unseen data lying outside the prior, even with latent code optimization at test time. Optimizing a randomly initialized generator, on the other hand, is prone to be trapped in a local minimum due to the complex energy landscape (path b). Whereas the pre-train prior could provide a good initialization in a forward pass, we propose to further optimize the parameters of the prior and the latent code according to the task-specific constraints at test time. We show in this work that the proposed framework can effectively break the barriers of pre-trained prior and generalize to the unseen data that is out of the prior domain (path c). Hence, our approach can generate results (ending point of path c) closest to the ground truth (star point on the real data manifold) compared to the other learning methods.

*Equal contribution
†Correspondence to Kui Jia <kuijia@scut.edu.cn>

The quality of the learned prior highly relies on the scale and diversities of the training examples. Yet, even with a large amount of data, the prior learned by the neural network may still be a crude approximation of the real data distribution (see Figure 1), making the network vulnerable to unseen data. This is particularly true with the 3D learning tasks, where the ground-truths are notoriously difficult to obtain, which greatly limits the scale of the training samples. Optimization-based approaches that leverage constraints from data, e.g. multi-view consistency, do not require any training to be usable. However, they are strict with the inputs and tend to fail on the sparsity of the data (e.g. single/sparse-view reconstruction) or the physical misalignment (e.g. unregistered/mismatched images).

To alleviate the generality issue of the learning-based approach while maintaining a friendly requirement for the inputs, we advocate a new 3D learning paradigm that connects the good ends of both learning-based and optimization-based approaches. In particular, we propose that the pre-trained data prior could obtain a maximum generality if it is *optimized*, rather than *fixed*, according to the data constraints at test time. Our approach shares a similar incentive with deep image priors [30], where high-quality images can be synthesized simply by optimizing an untrained and randomly initialized deep generator. However, unlike image synthesis, we show that optimizing a randomly initialized neural network often fails to achieve satisfactory results in 3D learning, especially in highly ill-posed configurations, such as sparse-view based 3D reconstruction.

Instead of fixing the priors or using random priors, we propose to jointly optimize the pre-trained shape prior and the latent code towards the input physical measurements at test time. Our observation is that though the learned prior cannot capture the full landscape of the real data distribution, it does provide a fairly good initialization for searching for the optimal solution in the entire embedding space (Figure 1). Further, by introducing the physically based optimization, the searching path could break the barrier of the pre-trained priors and converge at some point on the real prior which is more realistic but unreachable by only searching inside the learned priors (Figure 1). While it is possible that the optimization may lead to 3D shapes that do not look plausible, we propose that an l_2 regularization works surprisingly well in regularizing the searching space.

We materialize our idea using the implicit surface representation, as it is flexible to handle shapes with arbitrary topologies. We show that our proposed approach is a general 3D learning framework that supports a wide range of downstream applications, including shape modeling and reconstruction, with various forms of inputs. We also demonstrate that our framework can significantly improve the generality of the learning-based approach, even in the presence of highly sparse or collapsed observations, e.g. the sparse point clouds obtained from the 3D scanning, single or sparse views of the object of interest, etc. We verify the effectiveness of our approach in a variety of challenging tasks, including shape auto-encoding, sparse-view reconstruction and sparse point cloud reconstruction. Experimental results show that our approach is superior to the state-of-the-arts both quantitatively and qualitatively.

2. Related Work

Optimization-based Shape Reconstruction. Traditional image-based surface reconstruction methods, including PMVS [8] and COLMAP [26], *etc*., are mainly based on texture-rich and dense views for extracting multi-view correspondences. Since these approaches follow the exact data constraints, the reconstructed surface could be highly accurate. Nonetheless, they are also vulnerable to noisy input and collapsed observations which could interrupt the acquisition of pixel-wise correspondence across different views. In addition, they fail to generate plausible results in the presence of sparse views. The other line of research strives to reconstruct 3D surface from raw point clouds. The most representative ones include Poisson surface reconstruction [13], radius basis functions (RBF) [2], and moving least squares (MLS) [15] based approaches. The main idea of these methods is to fit either polygonal meshes or implicit functions to the input point cloud by optimizing a pre-defined energy objective. In contrast, our proposed method take advantage of both learning-based and optimization-based framework. Specifically, while we are able to faithfully reconstruct 3D surface with sparse observations, we can also achieve similar quality of reconstruction with the traditional stereo-based approach when observations are sufficient.

Learning-based Shape Modeling and Reconstruction. Recent years have witnessed great progress in introducing deep learning to 3D shape modeling and reconstruction. In particular, most of the previous works mainly rely on a retrieval based framework that fixes the parameters of the generator after training and retrieves the closest prior in the latent space via forward passing. It has been widely used in a wide range of 3D representations, including mesh [31, 32, 10, 24, 27], voxels [6, 29, 11], and implicit field [22, 4, 34, 28]. Though reasonable results can be obtained from these methods, they are vulnerable to disturbances in the input. Once the forward pass failed, one cannot further modify or optimize the results. To resolve this issue, recent works [1, 19] have proposed to optimize the latent code at test time. DeepSDF [25] presents the framework of auto-decoder where the shape prior is learned only with a decoder during training. The latent codes are optimized according to the input observations at test time, given a fixed pre-trained decoder. Despite that the reconstruction accuracy has been further improved by these approaches,

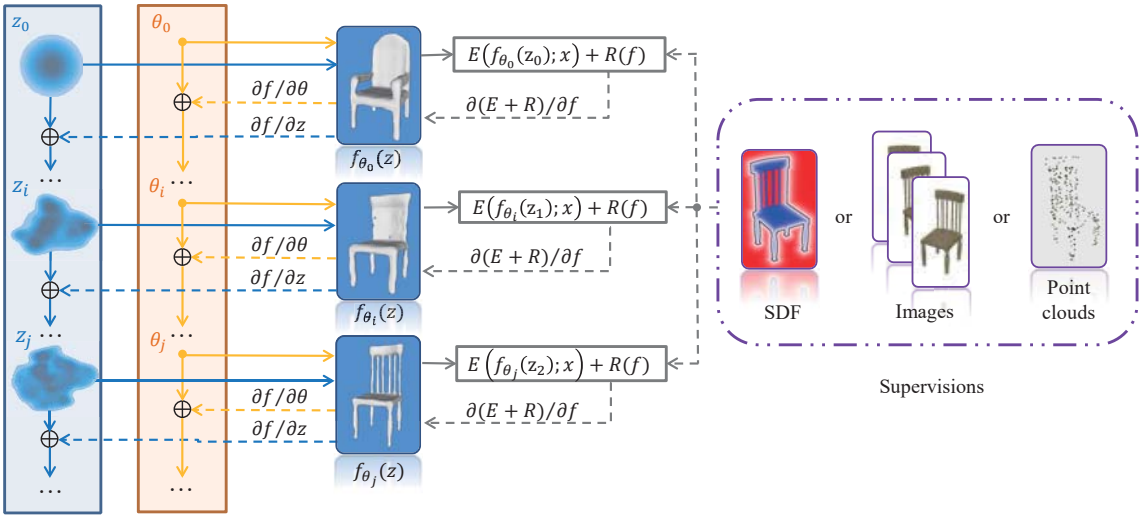

Figure 2: Illustration of our pipeline. Given specific supervisions (right) from the downstream applications, our goal is to optimize both the latent code z and the parameters θ of the pre-trained shape prior to generate a high-fidelity 3D result (left), which is represented by a neural implicit field. We perform iterative optimizations that jointly optimize z and θ in each iteration. Our optimization framework can gradually break the barrier of the pre-trained shape prior and converge to a faithful 3D shape that could lie out of the manifold spanned by training samples. We visualize the distribution of latent code in the leftmost using t-SNE technique [21] and the reconstructed results using marching cube [20] for better visualization.

they still have difficulty generalizing to unseen data as the pre-trained prior is limited to the domain spanned by the training samples. Recently, Williams et al. [33] propose to overfit a randomly initialized the neural network to an input point cloud. While surprisingly well results can be achieved in this setting, we show that it can hardly be applied to other challenging tasks, such as sparse-view surface reconstruction. In this paper, we propose a more general learning framework that strives to optimize both the pre-trained shape prior and latent code at test time. We show that it can significantly improve the generality and performance of the deep learned model in a wide range of highly ill-posed problems.

Combination of learning with data constraints. There have been a few preliminary explorations attempting to combine deep learning with optimization based on data constraints. In particular, [36] and [16] strive to reduce the searching space of a traditional optimization problem using a deep learned prior. They first encode the shape into latent space, and then optimize the latent code according to the photometric consistency constraint [16] or the bundle adjustment loss [36]. Though results that are more physically plausible can be achieved with these approaches, their performance is still limited to the quality of pre-trained prior and thus struggles to scale to unseen data. Another line of research aims to introduce data constraints as the supervision signal during training. For instance, the latest advances in differentiable rendering [17, 12] have been widely used in achieving unsupervised learning of single-view 3D mesh reconstruction. To leverage the flexibility of implicit representation, recent works [18, 23, 19, 35] have proposed new techniques to render implicit surfaces differentiably. These approaches succeed in training more powerful priors as there are ample resources of 2D images that can be directly used for training. However, they still do not resolve the generalization problem. Besides, since differentiable rendering techniques make 2D supervision possible, these methods [23, 35] can generate a single shape using dense views without learnt prior. In these cases, they can be considered as optimization-based methods. In contrast, our approach aims to incorporate the task-specific data constraints for optimizing shape priors and latent codes at test time, which can further boost the performance and generality of the previous approaches.

3. Methods

We interpret the implicit function learned by a deep generator as $f_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})$, which maps a query 3D point $\boldsymbol{x}$ and a latent code $\boldsymbol{z}$ to the target shape's approximate signed distance field. Our goal is to generate or reconstruct a faithful 3D surface $\mathcal{O}$ from the input physical observations, *e.g.* spatial signed distance field $\mathcal{S}$, sparse multi-view images $\mathcal{I}$, or point clouds $\mathcal{P}$, *etc.*, by leveraging the priors encoded in $f_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})$. Given the decoding model $f_{\boldsymbol{\theta}}$, the continuous surface associated with a latent code $\boldsymbol{z}$ is represented by the decision boundary of $f_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})$, and the shape can be

instantiated by isosurface extractions [20, 14]. Formally, a general 3D modeling problem can be formulated as follows:

$$\hat{\mathcal{O}}^* = \arg\min_{\hat{\mathcal{O}}} E(\hat{\mathcal{O}}; \mathcal{O}) + R(\hat{\mathcal{O}}), \tag{1}$$

where $E(\hat{\mathcal{O}}, \mathcal{O})$ is a task-specific energy term, and $R(\hat{\mathcal{O}})$ is a regularizer that encourages a plausible surface.

In deep image prior [30], the regularizer $R(\hat{\mathcal{O}})$ is realized using a randomly initialized and untrained neural network. However, unlike image synthesis, the 3D inverse problems are much harder, where merely depending on the prior brought by the structure of neural network is insufficient. Therefore, instead of random initialization, we leverage a pre-training to initialize $\boldsymbol{\theta}$ and $\boldsymbol{z}$ more appropriately. Hence, our goal is defined in the following form:

$$\boldsymbol{z}^*, \boldsymbol{\theta}^* = \arg\min_{\boldsymbol{z},\boldsymbol{\theta}} E(f_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}); \mathcal{O}) + R(f_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})), \tag{2}$$

where, instead of optimizing from scratch or a randomized neural network, we advocate to iteratively optimize the pre-trained priors, including the generator parameters $\boldsymbol{\theta}$ and the latent code $\boldsymbol{z}$, according to the data constraints $E(f_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}); \mathcal{O})$ (Figure 2). Further, the learned prior can be used as a strong regularizer $R(f_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}))$ to ensure a reasonable output. The formulation in Eq. (2) shows our general framework that combines the learning-based and optimization-based approaches. We will show in the next section how this formulation can be adapted to various applications.

4. Applications

We now show experimentally how the proposed approach works for diverse tasks on 3D modeling and reconstruction that take different input forms. Note that each application requires a pre-trained shape prior. Since the main focus of this work is not about how to obtain a stronger prior, we provide the details of our pre-training in the supp. material.

4.1. Shape Auto-Encoding

Auto-encoding 3D shapes play an important role in obtaining shape priors and a variety of downstream applications related to shape modeling and reconstruction. Since we implement our framework using implicit surface representation, our goal is to generate an implicit field as a faithful approximation of the input surface $\mathcal{S}$. We first convert the 3D locations to be queried into a signed distance field. The resulted field is composed of a set of pair $\{(\boldsymbol{p}_i, s_i)\}_{i=1}^n$, where the first element is the coordinates of the querying position in the space and the second element is its corresponding distance value. In particular, the reconstruction energy term in Eq. (2) is represented as

$$E(f_{\boldsymbol{\theta}}(\boldsymbol{z}); \mathcal{X}) = \sum_{i \in \{1,\dots,n\}} \|\hat{s}_i - s_i\|_1, \tag{3}$$

where s_i is the ground-truth distance; $\hat{\boldsymbol{s}}_i$ denotes the estimated signed distance value for the i^{th} point $\boldsymbol{p}_i$, predicted via the neural implicit field $f_{\boldsymbol{\theta}}(\boldsymbol{z}, \boldsymbol{p}_i)$. We apply the regularizer as that in Eq. (5), which will be discussed later in Section 4.2, namely $R(f_{\boldsymbol{\theta}}(\boldsymbol{z}))$, to ensure high-fidelity and reasonable results. This leads to a similar overall objective function as shown in Eq. (6).

4.2. Multi-view Reconstruction

Given a collection of multi-view images $\mathcal{I}$, together with object silhouette masks $\mathcal{M}$, the camera extrinsics $\mathcal{P}$ and intrinsics $\mathcal{K}$, the aim of multi-view reconstruction is to recover the underlying object surface from these partial observations of n views. To correlate the 3D surface with the 2D observations, we leverage the differentiable rendering technique such that the renderings of the generated surface are consistent with the input views. For more details on the differentiable rendering technique we used, please refer to the supp. material. Thereby, the energy term in Eq (2) is formulated as:

$$E(f_{\boldsymbol{\theta}}(\boldsymbol{z}); \mathcal{X}) = \sum_{i=1}^{n} (\|\hat{\boldsymbol{I}}_i - \boldsymbol{I}_i\|_1 + \lambda_c \cdot \mathcal{L}_c(\hat{\boldsymbol{M}}_i - \boldsymbol{M}_i)) \tag{4}$$

where $\mathcal{L}_c$ is the binary cross entropy, and λ_c is the weighted parameter. $\hat{\boldsymbol{I}}_i$ and $\hat{\boldsymbol{M}}_i$ denotes the estimated image and silhouette respectively for the i^{th} view. Specifically, the first term restrains only on the pixels inside the intersection of the given mask $\boldsymbol{M}_i$ and the predicted mask $\hat{\boldsymbol{M}}_i$, where the photometric RGB loss can be defined reasonably; while the second term applies to all the pixels to penalize mismatched object silhouettes.

In the presence of highly sparse views, the multi-view reconstruction task becomes a highly underdertermined problem. Hence, we further introduce additional regularizers on the neural network to ensure plausible results. For $\boldsymbol{z}$, we encourage the prior distribution of the latent code to be a zero-mean multivariate-Gaussian to encapsulate them into a compact shape manifold, preventing biased solutions. In addition, we would like to prevent $\boldsymbol{\theta}$ from moving too far away from the learned categorical prior. Through extensive experiments, we find that a simple l_2 norm on $\boldsymbol{\theta}$ works surprisingly well to strike a balance between flexibility and regularity. Formally, the regularizer term in Eq. (2) is defined as:

$$R(f_{\boldsymbol{\theta}}(\boldsymbol{z})) = \frac{1}{\sigma^2}\|\boldsymbol{z}\|_2 + \lambda_\theta \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2, \tag{5}$$

where λ_θ denotes the weighted parameter, and $\boldsymbol{\theta}_0$ denotes the parameters of $\boldsymbol{\theta}$ learned from the pre-training dataset. All together, the energy objective is formulated as:

$$\min_{\boldsymbol{\theta},\boldsymbol{z}} L(\mathcal{X}) = E(f_{\boldsymbol{\theta}}(\boldsymbol{z}); \mathcal{X}) + \lambda \cdot R(f_{\boldsymbol{\theta}}(\boldsymbol{z})), \tag{6}$$

where λ is the regularizer parameter. The overall default values are set as $\lambda = 0.5$, $\lambda_c = 0.5$, $\lambda_\theta = 0.1$, which works well in all our experiments.

4.3. Point Cloud Reconstruction

Our approach also supports reconstructing a complete 3D shape from the sparse 3D observation – point cloud $\mathcal{P}$. In this case, the input $\mathcal{X}$ is composed of a set of 3D points $\{\boldsymbol{p}_i\}_{i=1}^n$ with or without their corresponding normals $\{\boldsymbol{n}_i\}_{i=1}^n$. The goal is to reconstruct the continuous implicit field to represent the plausible object surface $\mathcal{O}$ that best fit the inputs. We hence can reformulate the energy term in Eq. (2) as:

$$E(f_{\boldsymbol{\theta}}(\boldsymbol{z}); \mathcal{X}) = \sum_{i \in \{1,\dots,n\}} (\|\hat{s}_i\|_1 + \lambda_n \cdot \|\hat{\boldsymbol{n}}_i - \boldsymbol{n}_i\|_2), \quad (7)$$

where λ_n is the weight for the normal regularizer; $\hat{s}_i = f_{\boldsymbol{\theta}}(\boldsymbol{z}, \boldsymbol{p}_i)$ and $\hat{\boldsymbol{n}}_i = \nabla_{\boldsymbol{p}} f_{\boldsymbol{\theta}}(\boldsymbol{z}, \boldsymbol{p}_i)$ are the estimated signed distance value and normal for the i^{th} point $\boldsymbol{p}_i$ respectively. Note that the normal term is optional depending on the availability of the normal data.

To encourage a smooth surface, apart from the multivariate-Gaussian prior for the latent space, we also include an Eikonal term [7], which regularizes the l_2-norm of the gradients $\nabla_{\boldsymbol{p}} f(\boldsymbol{z}, \boldsymbol{p}_i)$. The regularization term can be formulated as:

$$\begin{aligned} R(f_{\boldsymbol{\theta}}(\boldsymbol{z})) = \frac{1}{\sigma^2}\|\boldsymbol{z}\|_2 + \lambda_\theta \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \\ + \lambda_g \cdot \mathbb{E}_{\boldsymbol{p}}(\|\nabla_{\boldsymbol{p}} f_{\boldsymbol{\theta}}(\boldsymbol{z}, \boldsymbol{p}_i)\|_2 - 1)^2, \end{aligned} \quad (8)$$

where λ_θ and λ_g are the weights for their regularization terms. The Eikonal term is formulated as the expectation with respect to the probability distribution of $\boldsymbol{p}$. As it encourages the gradients $\nabla_{\boldsymbol{p}} f_\theta$ to be of unit-2 norm, f_θ will achieve minimum loss of Eq. (8) if f_θ vanishes on $\boldsymbol{p}$ and becomes a signed distance in Euclidean metric.

5. Experimental Results

Dataset. We adopt the category of chairs, lamps and cars in ShapeNet Core dataset [3] as our dataset, with $6778, 2318, 7497$ shapes respectively. Each mesh is normalized into a unit sphere during pre-processing. For the task of auto-encoding given input shape, we follow Park *et al*. [25] to construct the signed distance fields, each with $250,000$ spatial points and their values. For surface reconstruction based on sparse input RGB images, we use the rendered dataset from the Choy *et al*. [6] to adhere to the community standards [23, 22, 31, 32]. The dataset contains 24 images of resolution 64^2 and the viewpoints are sampled on the northern hemisphere of the object. We use 24 images and corresponding object masks per object for supervision on the pre-training stage to obtain a good prior, and 3 images from the testing set of the same resolution for testing. As for the task of point cloud based reconstruction, we sample $150,000$ points and their corresponding normals for training shape priors, but only use 300 points for evaluating our performance on sparse point cloud reconstruction.

Figure 3: Qualitative comparisons between shape auto-encoding results generated by different methods.

Evaluation metrics. For quantitative evaluations, we apply the most commonly used metrics of Chamfer Distance (CD) between uniformly sampled point clouds to measure the accuracy and completeness of the surface (the lower the better). We also adopt F-Score, measuring the completeness and precision of generated shapes (the higher the better). For shape auto-encoding, we further adopt the median of Chamfer Distance (the lower the better) following [25]. For point cloud reconstruction, we further use normal consistency [5] to measure the accuracy and completeness of the shape normals (the higher the better).

5.1. Shape Auto-Encoding

We compare the performance of shape auto-encoding with DeepSDF [25] in this section. We show the quantitative and qualitative results in Table 1 and Figure 3 respectively. As shown in Figure 3, our approach performs significantly better in recovering the fine details, such as the bumping details in the chair legs (1st column) and the thin rods on the chair back (5-th column). This performance leap become more prominent when the testing object deviates stronger from the training set. The quantitative results in Table 1 further support that our method performs much better across a range of different instances compared to DeepSDF [25].

5.2. Sparse Multi-view Reconstruction

Comparisons. We compare our proposed method with the state-of-the-art approaches including LSM [11], P2M++

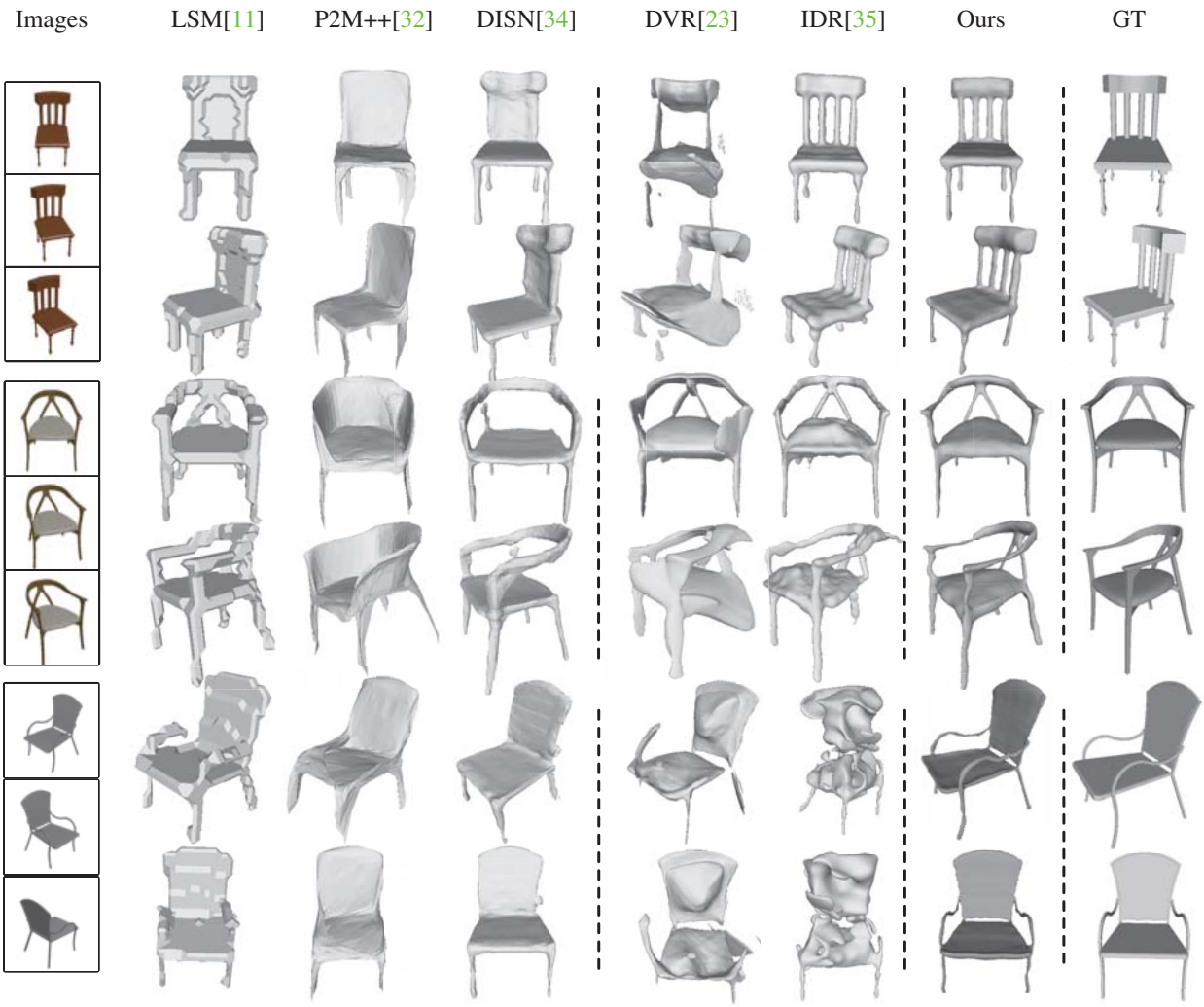

Figure 4: Qualitative comparisons among sparse multi-view reconstruction results generated by different methods. For each instance, the leftmost three images are the only given images, and the rightmost is the ground truth, dubbed as "GT". We show two different views for each generated object surface, including one from the view of the first image and one completely different view from the three images. We use dashed lines to distinguish the methods based on different concepts.

Methods	CD, mean (10^{-3})		CD, median (10^{-3})		F-score	
	Chair	Table	Chair	Table	Chair	Table
DeepSDF	0.21	0.42	0.08	0.07	88.23	82.34
Ours	**0.08**	**0.10**	**0.06**	**0.05**	**96.21**	**93.38**

Table 1: Comparative results of different methods for shape auto-encoding reconstruction. Performance is measured in terms the mean, median value of Chamfer distance and F-score over 50 instances. 10^{-3} refers to the magnitude.

[32], DISN [34], DVR [23] and IDR [35]. Specifically, LSM, P2M++, and DISN rely on retrieving the most likely shape priors via forward pass and are the representative works of voxel-, mesh- and implicit function-based approaches respectively. The other two methods can be considered as optimization-based methods here, which optimize a specific shape using data constraints. DVR [23] proposes a differentiable renderer for implicit field that enables unsupervised learning of 3D shape with 2D-to-3D consistency. IDR [35] achieves the state-of-the-arts for multi-view reconstruction. In particular, our approach employs the differentiable renderer from DVR in our network training. We randomly select 50 instances with 3 views for each object from the testing set per category. Results of these methods are obtained either by using their released codes (if available) or reproducing their methods (multi-view setting in DISN). In both cases, we report the best performance.

In Table 2, we compare different methods under the metrics of Chamfer and F-score. Our proposed method significantly outperforms all the alternative methods in both

Methods	CD, mean (10^{-3})		F-score	
	Chair	Lamp	Chair	Lamp
LSM [11]	7.36	6.32	27.43	25.89
P2M++ [32]	8.41	7.89	37.23	32.15
DISN [34]	2.75	15.29	52.47	26.03
Ours	**1.75**	**5.44**	**62.34**	**36.20**

Table 2: Comparative results of different methods for sparse view reconstruction. Performance is measured in terms Chamfer and F-score over 50 instances. 10^{-3} refers to the magnitude. We only show the comparative results with learning-based methods because optimization-based methods are unstable in some instances, which leads to terribly bad numerical results.

metrics. We demonstrate the visual comparison results in Figure 4, where we show the reconstructed shapes of several randomly selected instances. The retrieval based approaches, including LSM, P2M++, and DISN, can recover the rough shape and structure but struggle to capture fine-scale geometry details. This is primarily due to that they heavily rely on the pre-trained prior and have difficulty generalizing to the unseen data. The results from DVR [23] and IDR [35] in some cases perform well from the same views of input images (as shown in the first row of each instance), but appear significantly worse in the views without supervision. This is because that the training of DVR and IDR only resort to the specific data constraints, such as multi-view consistency or 2D-to-3D correspondence. As a result, their networks are prioritized to memorize the image-to-shape correspondence but lacks of a strong shape prior. This leads to corrupted results of DVR and IDR as shown in the third group of Figure 4. Since our approach leverages both the pre-trained shape prior and the novel optimization scheme, we are able to precisely reconstruct the intricate and thin structures, e.g, the thin chair legs and back structure, while ensuring a plausible shape. For more qualitative results, please refer to the supp. material.

Control Studies on Number of Views. We also conduct ablation study to evaluate the performance of our approach given different number of input views. In particular, we test our approach using 1, 2, 6 and 12 views. As can be seen in Figure 5, our approach can produce robust reconstruction with only a single view. In addition, with more views available, the quality of our reconstruction can be further improved. When 12 views are present, we can achieve similar quality of reconstruction with that of the stereo-based approach. It indicates that the learning-based approach can benefit a lot optimizing the pre-trained prior according to the data constraints. In Figure 6, we further compare our approach with DISN, which is specialized for singe-view reconstruction, and with IDR that excels at using dense multi-views. In comparisons, our approach can achieve similar or even better reconstructions.

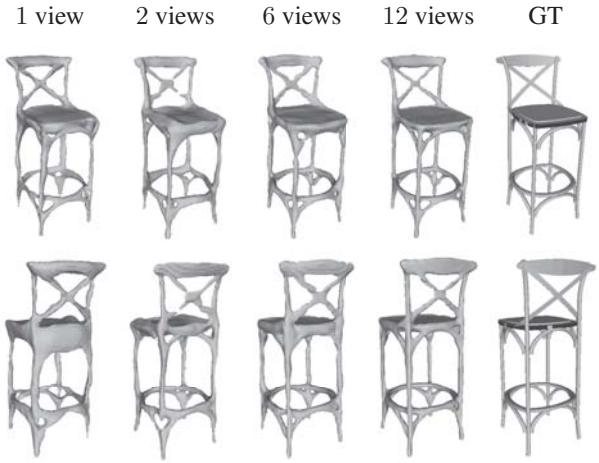

Figure 5: Example results of control studies on different number of views.

(a) Single View (b) Dense View

Figure 6: Qualitative comparisons among reconstructed results for the extreme cases of single view and dense views. We compare with the state-of-the-art method learning-based DISN [34] under single view setting while comparing with the state-of-the-art optimization-based method IDR[35] under dense view setting.

5.3. Point Cloud Reconstruction

We compare our approach with the state-of-the-art point reconstruction approaches: IFNet [5] and IGR [9]. We test all the approaches using a highly sparse point cloud consisting of 300 points. The qualitative and quantitative results can be found in Figure 7 and Table 3. Compared to IFNet and IGR, our approach can better reconstruct the intricate geometry details, such as the thin slats in the chair back (1st row), with quality close to the ground truth. In contrast, IGR fails to generate the thin chair legs (2nd row) while IFNet suffers from artifacts (1st and 3rd rows). In particular, we achieve these results by first searching for the instance shape code that best corresponds to the given point cloud, and then gradually updates the parameters of the implicit prior for the shape fitting. The initial latent code plays an important role for regularizing the subsequent optimization and provide an accurate initialization to optimize from.

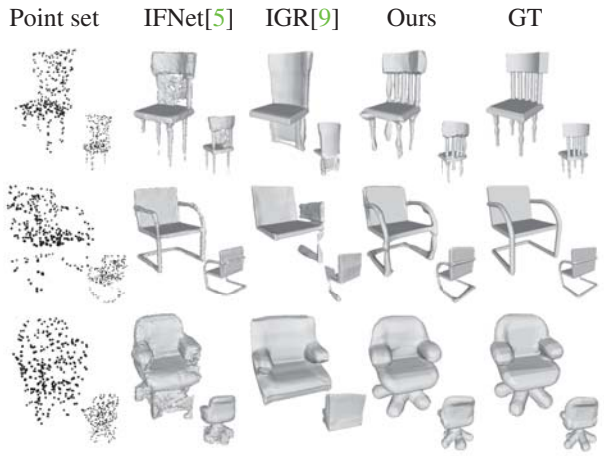

Figure 7: Qualitative comparisons between reconstruction results generated by different methods. The reconstruction results are based on the leftmost sparse point cloud.

Methods	CD, mean (10^{-3})		Normal-Consis.		F-score	
	Chair	Car	Chair	Car	Chair	Car
IFNet [5]	0.72	1.68	0.84	0.85	72.23	63.86
IGR [9]	1.55	0.71	0.75	0.89	60.32	73.39
Ours	**0.64**	**0.28**	**0.86**	**0.90**	**76.23**	**85.98**

Table 3: Comparative results of different methods for shape reconstruction from sparse point clouds. Performance is measured in terms of Chamfer distance, normal consistency and F-score over 50 instances. 10^{-3} refers to the magnitude.

5.4. Ablation Studies

In this section, we perform ablation studies to evaluate the efficacy of our proposed pipeline. All the following experiments are conducted in the context of multi-view reconstruction. Results are shown in Figure 8, including our proposed method, optimizing from random initialized network and optimizing only latent code. The reconstructions are based on the leftmost image (single view reconstruction), and then in turns our proposed method, the one without pretrained network parameters and the one without optimizing network parameters. We show two different views for each generated object surface, including one from the view of the image and one from a completely different view.

Optimize latent code only. One of the keys to our approach is optimizing both the latent code and the parameters of the network, making them adapt to the given observation. Though DeepSDF [25] has shown promising results by only optimizing latent code, we find in many cases, such as the 4th column in Figure 8, fails to faithfully reconstruct the geometric details, especially for unseen data. In contrast, our approach can achieve better results by jointly optimizing the latent code and the shape prior.

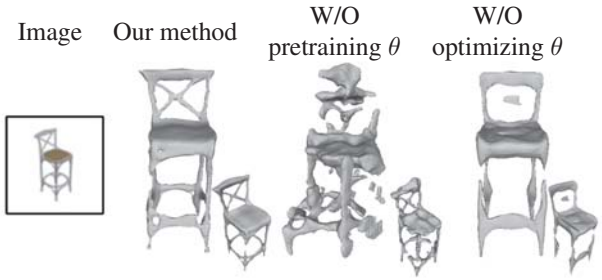

Figure 8: Example results of ablation study among our proposed method, optimizing both the latent code and the randomly initialized network, and optimizing the latent code only.

Optimizing from randomly initialized parameters. The other key for our approach is to pre-train the parameters of the network via an additional dataset to obtain a good initialization. As seen in Figure 8, optimizing a randomly initialized network (the 3rd column) fails to generate a plausible shape. Without the proposed pre-training, the subsequent optimization may deviate from a plausible searching path especially for highly ill-posed problems, such as single-view 3D reconstruction.

6. Conclusions and Discussions

We have presented a new learning framework for 3D modeling and reconstruction that profits from both the advantages of learning-based and optimization-based approaches. We have shown that by jointly optimizing the pre-trained prior and the latent code at test time, according to the data constraints, is a promising avenue to greatly improve the generality of a deep prior. To ensure the optimization would lead to reasonable result, we proposed that a simple l_2 regularization mechanism plays an important role in regularizing the searching space. Our experiments and evaluations have shown that our approach can generalize significantly better to unseen data compared to alternative approaches, especially in presence of sparse or highly collapsed inputs. Despite these promising directions, our method is currently more expensive than alternatives. It would be an interesting avenue to accelerate the optimization. In addition, we are still lacking a theoretical analysis of the working principle of our approach, which will be our next focus.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China (Grant No.: 61771201), the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (Grant No.: 2017ZT07X183), and the Guangdong RD key project of China (Grant No.: 2019B010155001).

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018. 2

[2] Jonathan C Carr, Richard K Beatson, Jon B Cherrie, Tim J Mitchell, W Richard Fright, Bruce C McCallum, and Tim R Evans. Reconstruction and representation of 3d objects with radial basis functions. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 67–76, 2001. 2

[3] 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. 5

[4] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 2

[5] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6970–6981, 2020. 5, 7, 8

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

[7] Michael G Crandall and Pierre-Louis Lions. Viscosity solutions of hamilton-jacobi equations. *Transactions of the American mathematical society*, 277(1):1–42, 1983. 5

[8] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multi-view stereopsis (pmvs). In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007. 2

[9] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 7, 8

[10] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–224, 2018. 2

[11] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. In *Advances in neural information processing systems*, pages 365–376, 2017. 2, 5, 6, 7

[12] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3907–3916, 2018. 3

[13] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. 2

[14] Jiabao Lei and Kui Jia. Analytic marching: An analytic meshing solution from deep implicit surface networks. In *International Conference on Machine Learning 2020 ICML-20*, 7 2020. 4

[15] David Levin. Mesh-independent surface interpolation. In *Geometric modeling for scientific visualization*, pages 37–49. Springer, 2004. 2

[16] Chen-Hsuan Lin, Oliver Wang, Bryan C Russell, Eli Shechtman, Vladimir G Kim, Matthew Fisher, and Simon Lucey. Photometric mesh optimization for video-aligned 3d object reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[17] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019. 3

[18] Shichen Liu, Shunsuke Saito, Weikai Chen, and Hao Li. Learning to infer implicit surfaces without 3d supervision. In *Advances in Neural Information Processing Systems*, pages 8295–8306, 2019. 3

[19] Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2019–2028, 2020. 2, 3

[20] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 3, 4

[21] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 3

[22] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 2, 5

[23] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 5, 6, 7

[24] Junyi Pan, Xiaoguang Han, Weikai Chen, Jiapeng Tang, and Kui Jia. Deep mesh reconstruction from single rgb images via topology modification networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9964–9973, 2019. 2

[25] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 2, 5, 8

[26] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 501–518. Springer, 2016. 2

[27] Jiapeng Tang, Xiaoguang Han, Junyi Pan, Kui Jia, and Xin Tong. A skeleton-bridged deep learning approach for generating meshes of complex topologies from single rgb images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[28] Jiapeng Tang, Xiaoguang Han, Mingkui Tan, Xin Tong, and Kui Jia. Skeletonnet: A topology-preserving solution for learning mesh reconstruction of object surfaces from rgb images. *arXiv preprint arXiv:2008.05742*, 2020. 2

[29] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2088–2096, 2017. 2

[30] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018. 2, 4

[31] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–67, 2018. 2, 5

[32] Chao Wen, Yinda Zhang, Zhuwen Li, and Yanwei Fu. Pixel2mesh++: Multi-view 3d mesh generation via deformation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1042–1051, 2019. 2, 5, 6, 7

[33] Francis Williams, Teseo Schneider, Claudio Silva, Denis Zorin, Joan Bruna, and Daniele Panozzo. Deep geometric prior for surface reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10130–10139, 2019. 3

[34] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 492–502. Curran Associates, Inc., 2019. 2, 6, 7

[35] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *NeurIPS*, 2020. 3, 6, 7

[36] Rui Zhu, Chaoyang Wang, Chen-Hsuan Lin, Ziyan Wang, and Simon Lucey. Object-centric photometric bundle adjustment with deep shape prior. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 894–902. IEEE, 2018. 3