

DeformNet: Free-Form Deformation Network for 3D Shape Reconstruction from a Single Image

Andrey Kuryenkov*, Jingwei Ji*, Animesh Garg, Viraj Mehta, JunYoung Gwak,
Christopher Choy, Silvio Savarese
Stanford Vision and Learning Lab

Abstract

3D reconstruction from a single image is a key problem in multiple applications ranging from robotic manipulation to augmented reality. Prior methods have tackled this problem through generative models which predict 3D reconstructions as voxels or point clouds. However, these methods can be computationally expensive and miss fine details. We introduce a new differentiable layer for 3D data deformation and use it in DEFORMNET to learn a model for 3D reconstruction-through-deformation. DEFORMNET takes an image input, searches the nearest shape template from a database, and deforms the template to match the query image. We evaluate our approach on the ShapeNet dataset and show that - (a) the Free-Form Deformation layer is a powerful new building block for Deep Learning models that manipulate 3D data (b) DEFORMNET uses this FFD layer combined with shape retrieval for smooth and detail-preserving 3D reconstruction of qualitatively plausible point clouds with respect to a single query image (c) compared to other state-of-the-art 3D reconstruction methods, DEFORMNET quantitatively matches or outperforms their benchmarks by significant margins. For more information, visit: <https://deformnet-site.github.io/DeformNet-website/>.

1 Introduction

This paper studies the structured prediction problem of regressing unordered point sets with implicit and often ambiguous input spaces. A concrete instance which embodies this type of problem is 3D object geometry reconstruction (3DR) from single-view images for partial shape guidance [9]. The ambiguity arises from the fact that 3D-to-2D mapping is not invertible and large portions of the object features are typically occluded. 3DR is a pivotal learning problem in visual understanding, with numerous applications across domains. For instance, an intelligent robot requires a 3D model of the object instance to reason about manipulation. Similarly, in augmented reality recognizing the 3D shapes of often unseen objects in the world is necessary for both correct rendering and interaction.

3DR has been explored in a large body of extant work in computer vision, for problems such as structure from motion [10, 16] or multiview stereo [1, 7, 11, 12, 14, 18, 22] and at times even with single view images [6]. Ingenious work on “Shape-from-X” has utilized priors on natural images to infer geometric features, with “X” being shading, texture, specularity, shadow and so on [2, 17, 23, 28, 39]. Most of the aforementioned methods require carefully constructed features, a problem that is addressed by data-driven methods relying on large-scale 3D object datasets [3, 35].

Data-driven methods learn implicit priors for various object recognition tasks such as shape completion and 3D reconstruction. Broadly, these methods use the prior knowledge in two ways: (i) image-based shape retrieval that focuses on algorithm design to find the nearest shape in database for the query image [3, 26, 36], and (ii) deep generative models which operate directly on the query image and generate a 3D reconstruction as output, matching the shape distribution but resulting in different shape instances than in the database [5, 8, 15, 25, 32, 33, 37].

*Equal Contribution

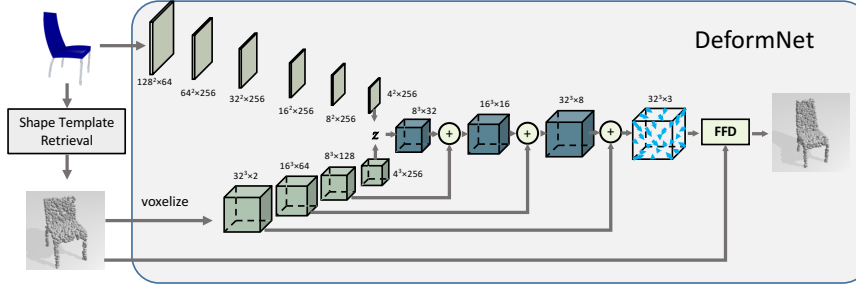


Figure 1: Our framework and DeformNet architecture. ‘+’ denotes stacking activations from image encoder, voxel encoder and voxel decoder. The output of decoder is the prediction on the offsets of control points, which decides the free-form deformation of input shape on the next step.

We note that the majority of recent methods for 3DR resort to either direct volumetric representation (aka voxels) or meshes from multi-view images as their shape representation. While intuitive, these representations can be both computationally inefficient and ineffective in capturing the natural invariance of 3D shapes under geometric transformations. A recent method by Fan et al. uses Point Set Generation Network (PSGN) to alleviate these problems. As they note, a point cloud is a simple, uniform structure that is relatively easier to learn than voxels, as it does not have to encode multiple primitives or combinatorial connectivity patterns. Additionally, point clouds are computationally superior to voxels since they do not require 3D convolutions and are amenable to direct manipulation when it comes to shape deformation and transformation. However, though direct prediction of fixed size point clouds improves 3DR performance, giving up on connectivity can result in a lack of fine shape features since loss functions are focused on overall reconstruction.

At the same time, Spatial Transformer Networks (STN) have presented an appealing method to learn geometric transformations in 2D images [20]. STNs are a modular, differentiable and dynamic upgrade to pooling operations that learn to zoom in and eliminate background clutter, thereby “standardizing” the input. However, they have primarily been studied in the context of discrete grids in image inputs to facilitate image classification.

Inspired the ideas from PSG and STN, we propose DEFORMNET- a model that extends STN style geometric operations to 3D using the notion of Free-Form Deformations (FFD). When used in conjunction with a point cloud representation, this method not only benefits from computational efficiency but also can preserve fine details in shapes since it implicitly preserves connectivity in structures. DEFORMNET uses a single image input to first perform shape retrieval from an object dataset using a learned image-to-shape embedding, and then deforms the point cloud representation of the retrieved template using the FFD layer in an encoder-decoder style network architecture as depicted in Figure 1. DEFORMNET intuitively builds on the strengths of Shape Retrieval, PSG, and STNs while compensating for their shortcomings. We implement FFD as a differentiable layer for end-to-end training along with point set correspondence based loss functions Chamfer Distance and Earth-Mover’s Distance, as in [8].

To summarize, the main contributions of this paper are:

- Introduction of Free Form Deformation as a differentiable layer to be used as a new building block that enables 3D data manipulation.
- A novel framework, DEFORMNET based on FFD layer to achieve smooth geometric deformations in point clouds for 3D Reconstruction.
- Evaluation of DEFORMNET on rendered images achieves state-of-the-art performance in comparison to both Point Prediction Methods such as PSG [8] and Generative Models such as 3D-R2N2 [5].

2 Related Work

Generative Models for 3D Reconstruction. Recently, generative models for 3D reconstruction have produced state-of-the-art results. One approach is targeting voxel reconstruction through a 3D voxel neural network. [5] proposed a 3D recurrent neural network (3D-R2N2) based on long-short-term memory (LSTM) to infer 3D volumetric shape from a single view or multiple views. Girdhar *et al.* [13] proposed a TL-embedding network which embeds image and shape together for single

view 3D volumetric shape generation. Wu *et al.* [34] proposed a 3D VAE-GAN which brings the two popular generative models together in volumetric shape generation and reconstruction. There are also representative advances in unsupervised/weakly supervised 3D learning for reconstruction. [37] proposed a perspective transformer net for reconstruction from a single image which only uses images contour as supervision. [25] proposed a conditional generative network for unsupervised reconstruction from images.

All of the above neural network based 3D reconstruction methods use voxel representations, which requires a large amount of memory and is inefficient due to the small space usage of generic 3D shapes. Fan *et al.* [8] proposed an alternative approach with neural networks that output unordered 3D point sets for 3D reconstruction. In our work, we combine both voxel and point set representations by using a 3D convolutional neural network to generate a deformation that is applied to 3D points sets for output which can preserve fine detail in the template shape.

Spatial Transformer Networks Conventional convolutional neural networks lack the ability to warp or select a patch from an image that is relevant to the target task, which leads to an unnecessarily larger and deeper network for larger images. To alleviate this issue, Jadenberg *et al.* [20] proposed the Spatial Transformer Networks (STN) that can apply geometric transformations to an input or activations such as Affine Transformation or Thin-Plane-Spline to extract a patch that is relevant to the task. Kanazawa *et al.* [21] proposed a WarpNet, an unsupervised method for deforming an image using a Spatial Transformer Network. This work is similar in spirit to the WarpNet in that the neural network generates deformation parameters and the loss is defined using the deviation of the deformed input to the ground truth. However, unlike the WarpNet, we condition the network on two inputs (the reference image and a template shape) and use it for 3D shape deformation and reconstruction.

3D Shape Deformation Recently, Yumer and Mitra [38] proposed a 3D Convolutional Neural Network that generates a deformation field as an output for 3D mesh editing given a user input. We employ the same deformation representation, Free-Form Deformation (FFD) to generate 3D deformation field.

The primary difference between our approach and that of Yumer *et al.* are that they supervise the network using precomputed deformation offsets while we propose an end-to-end trainable network which only requires the target shape as the sole source of supervision. We accomplish this by computing distance between a set of points and minimizing it with respect to the deformation field (Sec. 3.4). In addition, unlike Yumer *et al.*, we focus on the 3D reconstruction given an image input, rather than 3D editing.

Huang *et al.* [19] also proposed an approach to 3D reconstruction through template retrieval and deformation, which relies on jointly segmenting the 2D image and 3D templates and creating the 3D reconstruction from deformed parts of the segmented templates. Their approach differs from ours in that it relies on having noisy segmentation of the shapes in their database, in that it performs reconstruction by segmenting both the input image and templates and deforming template parts to fit with the parts of the segmented image via direct optimization, and in that it takes significantly longer to run due to the optimizations it employs.

To summarize, voxelized representations suffer from memory inefficiency and difficulty in generating fine details, previous deformation networks require the limiting supervision of point displacements, and previous work on reconstruction through template deformation relied on noisy segmentation in the shapes database. Purely generative models don't leverage the high quality and broad availability of shape databases. Our paper finds a balance between the flexibility of a generative approach and the output quality of a database-focused approach. It does this without requiring any supervision beyond that of the desired shape. This addresses the major problems with the current best methods.

3 DEFORMNET

In this section, we propose the DEFORMNET framework that generates a 3D shape reconstruction from a single image. Unlike recent single-view 3D reconstruction works that use Convolutional Neural Networks, we do not use voxelized output [5, 13, 34], which cannot recover fine details due to the coarse resolution of the voxel grid, and do not directly generate a point cloud from scratch [8]. Instead, we propose an end-to-end network that deforms a template 3D shape to match an input image and preserves the topology of the template shape and train it using the target shape only. This can be seen as combining the key insights from prior work that used Free-Form Deformation in a deep learning model [38] and the Chamfer distance for the objective function [8].

In the following subsections, we first introduce metric learning based shape template retrieval, which finds the most topologically similar 3D shape template for a given image (Sec. 3.1). Then, we introduce DEFORMNET, which takes a set of templates that we use for deformation and a query image and learns to output the deformation field end-to-end given the target shape (Sec. 3.3). Lastly, we present two objective functions that measure the deviation of the deformed shape from the target shape by finding correspondences on-the-fly (Sec. 3.4). See Figure 1 for illustration.

3.1 Shape Template Retrieval

To make use of the high-quality 3D CAD models in the existing database, the first step of the framework is to retrieve shape templates that have a similar topology to the object in a query image. For this, we use metric learning to learn an embedding that preserves topological similarity between shapes. Specifically, we first define the metric space by a set of constraints that force dissimilar shapes to be at least a margin father away than the distance between similar objects.

$$d(F(x_i; \theta), F(x_j; \theta)) + \Delta < d(F(x_i; \theta), F(x_k; \theta)) \text{ where } i, j \in \mathcal{C}_n, k \in \mathcal{C}_m, m \neq n \quad (1)$$

where $\Delta \in \mathbb{R}$ is a margin, \mathcal{C}_n indicates a set containing all elements in the n -th class and $d(\cdot, \cdot)$ is an arbitrary distance function. We use a parametric metric space representation using a neural network $F(\cdot; \theta)$ where θ denote the parameters in the neural network. The above constraints can be converted to a loss function which forms a triplet loss [29]. By minimizing the loss function with respect to θ , we can generate a feature representation that preserves perceptual similarity in a metric space where the distance operation is meaningful. However, due to difficulty in mining hard negatives and inefficiency of not reusing features in a batch, the naive metric learning approach only yields a moderate result [24, 29]. Instead, we use the smoothed version of the triplet loss that reuses all features in a batch for efficient hard negative mining [24] and allows fast and effective training.

$$J = \frac{1}{2|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \left[\log \left(\sum_{k \in \mathcal{N}_i} \exp\{\Delta - d_{i,k}\} + \sum_{l \in \mathcal{N}_j} \exp\{\Delta - d_{j,l}\} \right) + d_{i,j} \right]_+^2 \quad (2)$$

where $d_{i,j}$ indicates $d(F(x_i; \theta), F(x_j; \theta))$ and \mathcal{N}_i denotes a set of shapes that belong to different categories from the category of i .

Specifically, we used a 2D CNN to implement $F(x_i; \theta) \in \mathbb{R}^D$ and x_i denotes rendering of the i -th shape. We define the positive pairs \mathcal{P} to be renderings of the same shape from different perspectives and negatives \mathcal{N} to be renderings from different shapes. We define the similarity to be the inverse of the distance in the metric space and retrieve K -nearest neighbors from a query image and use the shapes that generated the images for the next stage.

3.2 DEFORMNET: Model

Given a reference image and a shape template that closely matches the object in the input image from the Shape Retrieval stage (Sec. 3.1), we want to generate the parameters of a deformation which transforms the shape template into the shape in the reference image. Unlike conventional CNN for reconstruction, DEFORMNET takes two different modalities as inputs and thus has two CNNs in these respective domains to encode the inputs. First, for the 2D image, we used 2D CNN for an image encoder $E_I(I) \in \mathbb{R}^{F_I}$ and for the 3D shape input, we voxelized sampled points on the surface to generate dense point cloud and voxelized to feed into a 3D CNN encoder $E_s(S) \in \mathbb{R}^{F_s}$. We simply combined the information from two modes by stacking the final fully connected layer activation to the last 3D CNN activation along the channel so that we preserve the spatial information from the 3D shape template. The combined information from both encoders contains all information that we need to know from both inputs and thus call the latent variable $z \in \mathbb{R}^{F_I+F_s}$. z is then fed into a 3D decoder $D(z)$, a 3D Deconvolutional Neural Network, to expand the spatial dimension of the output. Since the 3D encoder loses spatial details as we project the activations from 3D convolution layers to a coarser voxel grid, we use a 3D U-Net structure which is an extension of the 2D U-Net proposed in [38] to recover the details in the output. The U-net has an hourglass shape and skip connections between the 3D encoder and 3D decoder that crosses the latent variable (Fig. 1).

The final output of the decoder is a vector field $V = \{v_i\}_{i=1, \dots, N^3}$, $v \in \mathbb{R}^3$ which is used as the offset for the N^3 control points in the Free-Form Deformation Layer (Sec. 3.3). Each offset v_i , represents x, y, z offset of the corresponding control point. These values are used to then compute the deformed point cloud output, which is the final output of the network, and which can then be compared against the ground truth point cloud of the input image shape.

3.3 Free-Form Deformation Layer

Free-Form Deformation (FFD) [30] is the 3D extension of a Bezier curve form, which has been widely used for shape deformation. Since it is defined on a 3D grid, FFD fits with 3D convolutional neural network and has been used for generating 3D deformation using a neural network before [38]. The DEFORMNET learns the FFD for every input image - shape template pair, and the predicted FFD on shape template will generate the final deformed shape. In this paper, we mainly focus on manipulating and evaluating reconstructions with point clouds, though the learned FFD could also be applied to other formats.

FFD is formally formulated as following. Let $p = (u, v, w)$ be the normalized point coordinate in the grid and Δ_{ijk} be the 3D deformation offset at the grid control point $p_{ijk} = (i, j, k)$. Then, the point p after deformation is defined as

$$p' = \sum_{i=0}^l \sum_{j=0}^m \sum_{k=0}^n (p_{ijk} + \Delta_{ijk}) B_{l,i}(u) B_{m,j}(v) B_{n,k}(w) \quad (3)$$

where $B_{n,m}(x) = \binom{n}{m} (1-x)^{n-m} x^m$ is a binomial function, and l, m, n are sizes of the control point grid. The interpolation will mix the displacement of all control points around each data point, generating a smooth deformed output. The 3D decoder outputs deformation field V . Note that p' is differentiable with respect to v_{ijk} , which guarantees that the backward propagated gradients could flow from the objective function on the top of FFD, making FFD learnable.

3.4 Objective Functions

To make the DEFORMNET end-to-end trainable, we need to define a loss function that optimizes the target task: deforming a template shape to match a target shape. Ideally, the function should measure difference between deformed shape and a template shape and should return 0 if and only if the deformed shape overlaps the template exactly. However, 3D shapes are defined by vertices and faces whose accurate topological similarity is difficult to measure. So, rather than measuring the topological similarity, we sample points on the surface of a shape and use the set of points (point cloud) as a surrogate for a shape. Point clouds are easy to manipulate due to their simplicity and efficiency, and we follow [8] by using distance functions on point clouds as the loss function. If the network generates the correct deformation field V that minimizes the point cloud distances, then the deformed point cloud will match the target point cloud accurately as well. We explore two point cloud distance functions: Earth Mover's distance (EMD) and Chamfer distance (CD) [27]. Both distance measures are based on point-wise L-2 distance.

Earth Mover's distance: The EMD is defined as the minimum of sum of distances between a point in one set and a point in another set over all possible permutation of correspondences. To find the minimum, the EMD implicitly solves bipartite matching problem. More formally, given two sets of points S_1, S_2 , the EMD is defined as

$$d_{EMD}(S_1, S_2) = \min_{\phi: S_1 \rightarrow S_2} \sum_{p \in S_1} \|p - \phi(p)\|_2 \quad (4)$$

where ϕ is some bijection from S_1 to S_2 .

Chamfer distance: CD is computationally easier than EMD, since it uses sub-optimal matching to determine the pair-wise relation. For each point in one point set, CD simply treat the nearest neighbor in the other set as the image of this point. Formally, CD is defined as

$$d_{CD}(S_1, S_2) = \sum_{p_1 \in S_1} \min_{p_2 \in S_2} \|p_1 - p_2\|_2^2 + \sum_{p_2 \in S_2} \min_{p_1 \in S_1} \|p_1 - p_2\|_2^2 \quad (5)$$

We use two forms of regularization in addition to the distance function, the first being L1 loss over all point cloud offsets to force the network to deform the template as little as possible, and the second being L2 loss over the difference between neighboring control point offsets to promote smooth deformation.

3.5 Implementation Details

For shape retrieval, we use the multiview rendered images of shape from [5] to train the network and used the GoogLeNet model [31]. At training and testing time, the template shape input is chosen

at random from the 5 most similar shapes from the same category. For EMD, we precomputed ground truth correspondences between each input and its 5 most similar shapes using the Hungarian Algorithm; thus it should be noted that the correspondence is based on the undeformed template, which is an approximation of true EMD. We evaluated using true EMD loss and found it to not work better, and as discussed in the appendix focused on using Chamfer distance as the loss for training. For Chamfer distance, we use the output of the network directly to compute the distance.

We used TensorFlow to implement the networks and used Adam optimizer with momentum term of 0.95. The model is trained with an initial learning rate of $5e-4$ and goes down to $5e-5$ after 20k iterations. After selection based on performance, we choose a batch size of 16 for training. We use leaky ReLU as an activation function. Note that to make use of EMD, we resample the point clouds to normalize the number of points. We train on point clouds with 1024 points. For deformation, we set $N=4$ as the number of control points in each dimension that points are computed with (so each point in the deformed output point cloud is a function of $N^3 = 64$ control points). We use $\lambda = 0.05$ for regularization on control point offsets.

4 Experimental Evaluation

4.1 Experimental Setup

We train and evaluate our models on the ShapeNet database [4], which contains a large quantity of manually created and cleaned 3D CAD models. Specifically, we select 5 representative categories to study on: chair, car, airplane, bench and sofa, following Gwak *et al.* [15]. The images for training and testing are rendered in various angles to provide synthetic training data for the model. In total, 22,324 shape models are covered, where training/testing split is 80%/20%. The 3D CAD objects are originally stored as meshes, so we enriched the dataset via resampling the meshes into point clouds and voxelizing them into voxels.

4.2 3D Shape Reconstruction from RGB Images

We compare our work with point set generation network (PSGN) [8], which is the state-of-the-art in deep learning based 3D reconstruction from a single image. PSGN chooses point clouds as the 3D representation, which allows manipulation including geometric transformations and deformations. Also, point clouds can in principle contain more information than voxel representations due to the latter’s discrete nature, and point clouds are easy to convert into voxels whereas the other way around will be tricky. Therefore we also target point clouds, though the learned free-form deformation can be applied to both voxel, point cloud, and mesh.

On point clouds, PSGN proposed two metrics for training and evaluating the reconstruction - CD and EMD. To have a fair comparison, we use the same point-set based metrics and follow their experiment setups. We train and test with relatively sparse point clouds with 1024 points, though as will be demonstrated in section 4.6 our model can be applied to dense point clouds despite being trained with sparse ones. To have comparable scaling of distances, during evaluation point clouds are bounded in a hemisphere with a radius of unit 1 and are aligned to their ground plane. Unit 1 is defined as 1/10 of the length of the 3D grid as done in PSGN. We train our models on all five categories with only the CD the loss (as it was found that EMD provided no benefit over CD), and provide both CD and EMD metrics on the test set alongside the same metrics for the newest trained model released by PSGN.

Here we report the per category comparison on both CD and EMD metrics in Table 1. In PSGN, they include the mean value of point-set based metrics from 3D-R2N2 [5], thus we also list them here. On CD metrics, we outperform PSGN and 3D-R2N2 on all categories; on EMD, PSGN achieves better values on the car and sofa categories, but our model performs significantly better in the other three categories as well as the average value. This indicates PSGN has good performance on rotund and

Table 1: Comparison with point set generation network and 3D-R2N2 on point-set based metrics. ‘Rtrv’ is the loss when just using one of the top 5 templates at random, without deformation. ‘No Rtrv’ is our model trained without shape retrieval, to deform a random in-category shape. The numbers are the average point-wise distances. (Lower is better)

Category	CD					EMD				
	Ours	Rtrv	No Rtrv	PSGN[8]	3D-R2N2[5]	Ours	Rtrv	No Rtrv	PSGN[8]	3D-R2N2[5]
airplane	0.10	0.15	0.20	0.14	-	0.56	0.64	0.74	1.15	-
bench	0.10	0.24	0.21	0.21	-	0.55	0.64	0.68	0.98	-
car	0.09	0.14	0.13	0.11	-	0.52	0.63	0.63	0.38	-
chair	0.13	0.19	0.27	0.33	-	0.51	0.62	0.70	0.77	-
sofa	0.21	0.30	0.37	0.23	-	0.77	0.83	0.84	0.60	-
mean	0.13	0.20	0.24	0.20	0.71	0.58	0.67	0.72	0.78	1.02

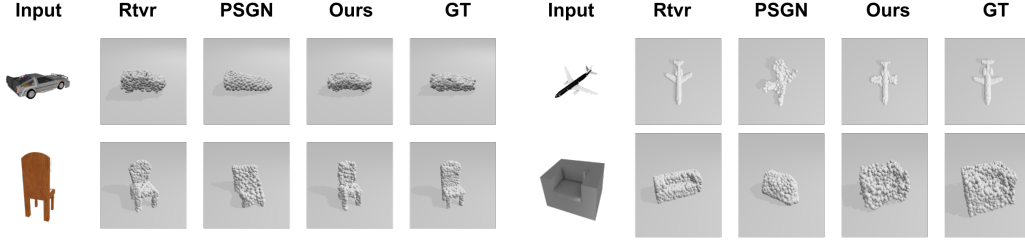


Figure 2: Visual comparison of different approaches. Left to right: input image, retrieved shape, point set generation network output, the output reconstruction trained with CD on our full model, ground truth shape. The examples are hand-picked from 4 categories.

less detailed objects, while our strength is on objects with more fine-grained details such as chairs and airplanes. Both our work and PSGN focus on reconstruction on point clouds, where the point-set based metrics are straightforward and intuitive.

To contrast the two methods clearly, we show the visual comparison in Figure 2. In general, our reconstruction can recover the main features from the object, even when the input template is not ideal for the image. This can be attributed to the combined benefits of shape retrieval and deformation - shape retrieval alone provides very good complete 3D shape templates without missing features, and deformation is able to preserve all of the template’s main features while tweaking them to more closely match their shape in the image input.

4.3 Ablation Analysis

Sensitivity to input shape template. Image-based shape retrieval could provide reasonable CAD model as template to start deforming with. We therefore also provide analysis on the degree to which DEFORMNET relies on having a good shape template.

To do this analysis, we compute the statistics on test set of all categories: for each group of input template (In), output (Out), ground truth shapes, we compute the tuple $(d_{CD}(\text{In}, \text{GT}), d_{CD}(\text{Out}, \text{GT}))$. Figure 3 shows the scattering plot of 516 random groups. Most of them lie closely to the horizontal axis, showing that DEFORMNET is not very sensitive to the input template’s quality, which indicates the robustness of DEFORMNET. We also report the CD and EMD with randomly picked shape template as input (without shape retrieval) in 2, averaged on all categories. Note that DEFORMNET without shape template retrieval still outperforms PSGN on average.

Joint skip connections. The joint skip connections stacking activations from 3D encoder and 3D decoder serve as information conveyors on the same level of spatial resolution. To verify its importance, we trained a network without the joint skip connections and measured its performance in CD and EMD metrics as shown in Table 2. Since the joint skip connections feed sharp information from the shape directly to deeper layers in the decoder, the reconstruction is more accurate in terms of CD loss.

Regularizer. To understand the functionality of the deformation regularizer, we have trained and tested a network without the regularization term in the objective function. Quantitatively, the model without regularization performs slightly better than the model with it, as shown in Table 2. However, the regularizer encourages more conservative and smooth deformation. For mesh reconstruction, this lowers the chance for faces to cross during deformation. As illustrated in the mesh reconstruction experiment in Figure 6, the regularizer enforces the consistency among offsets of neighboring control points, which is not guaranteed in the model without the regularizer.

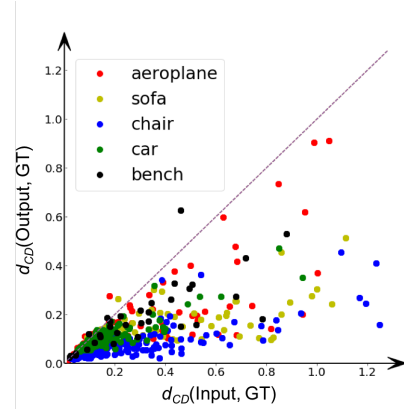


Figure 3: Sensitivity to shape template choice. The horizontal axis is the CD between ground truth and a random template input, and the vertical axis is the CD between ground truth and the output from DEFORMNET.

	CD	EMD
w/o skip	0.183	0.563
w/o reg	0.125	0.582
full	0.127	0.585

Table 2: Ablation analysis

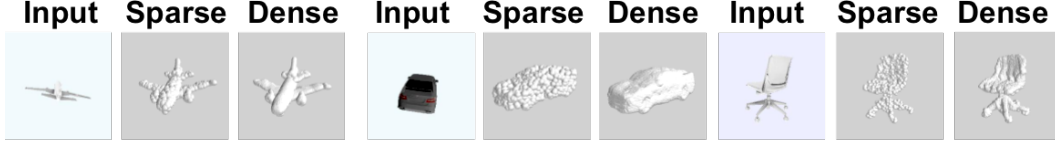


Figure 4: Reconstruction on sparse and dense point clouds. Sparse point cloud have 1024 points, dense point clouds have 16384 points. The model is trained on sparse point clouds only.

Point cloud density. One advantage of FFD is that the same deformation can be applied to any number of points in the control points grid, which bridges the gap between low and high resolution. To have a fair comparison with PSGN, we also trained on sparse point clouds with 1024 points, but our trained model could be directly used on deforming and reconstructing dense point clouds. Figure 4 shows the comparison between reconstructed sparse and dense point clouds, using the model trained on sparse point clouds. In general, network deforms dense point clouds similarly to sparse ones.

4.4 3D Reconstruction with Real Images



Figure 5: Real image reconstruction. On the first three test examples, DEFORMNET generates reasonable reconstruction, while the last one fails.

We also tested our model on real world images and visualization results are shown in Figure 5, including a failure case. Since we trained our model on synthesized images with single-colored backgrounds, we segmented real images as the input into network as done in PSGN. Our network successfully infers the 3D shape in some cases, but fails on some others. One solution could be domain adaptation and transfer learning, to bridge the gap between rendered and real world images, which we leave to future work.

4.5 Mesh Reconstruction

Free-form deformation has been widely used on deforming mesh objects, and with FFD layer in a deep neural network, the learnable manipulation and reconstruction on mesh become possible. As analyzed in section 4.3, the regularized model refrains from drastic changes in a local patch, which gives out mesh reconstruction with plausible quality. Figure 6 shows an example of mesh reconstruction on a chair.

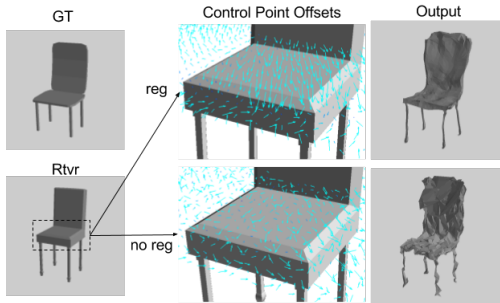


Figure 6: Comparison between the behavior of DeformNet with and without regularization during training. The ground truth and retrieved model are similar except for the thickness of the seat and its height. The arrows shown in the zoomed-in figures are the offsets of control points. Color denote the offset's magnitude. The regularized model learns to consistently squeeze the seat uniformly whereas the unregularized model displaces the control points less smoothly, which results in the difference in the output mesh reconstruction.

5 Conclusions and Outlook

This paper examines the structured prediction problem of regressing 3D point clouds based on image input to solve 3D Reconstruction with a single image. We note that existing methods with volumetric representations fall short on computational efficiency. This paper leverages the Point cloud-based representation coupled with a 3D generalization of the Spatial Transformer using Free-Form Deformation to achieve state of the art results on reconstruction with ShapeNet. We introduce Free Form Deformation as a differentiable layer to enable 3D data manipulation. This can have wider implications beyond 3D reconstruction, such as in point-cloud processing, and learning to perform grasping on unseen objects.

References

- [1] Sameer Agarwal, Noah Snavely, Ian Simon, Steven M Seitz, and Richard Szeliski. “Building rome in a day”. In: *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 72–79 (cit. on p. 1).
- [2] Jonathan T Barron and Jitendra Malik. “Shape, Illumination, and Reflectance from Shading”. In: *TPAMI* (2015) (cit. on p. 1).
- [3] 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* abs/1512.03012 (2015) (cit. on p. 1).
- [4] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing 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*. Tech. rep. Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015 (cit. on p. 6).
- [5] 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*. Springer, 2016, pp. 628–644 (cit. on pp. 1–3, 5, 6).
- [6] A. Criminisi, I. D. Reid, and A. Zisserman. “Single View Metrology”. In: *International Journal of Computer Vision* 40.2 (2000), pp. 123–148 (cit. on p. 1).
- [7] Jakob Engel, Thomas Schöps, and Daniel Cremers. “LSD-SLAM: Large-scale direct monocular SLAM”. In: *Computer Vision—ECCV 2014*. Springer, 2014 (cit. on p. 1).
- [8] Haoqiang Fan, Hao Su, and Leonidas Guibas. “A Point Set Generation Network for 3D Object Reconstruction from a Single Image”. In: *arXiv preprint arXiv:1612.00603* (2016) (cit. on pp. 1–3, 5, 6).
- [9] Michael Firman, Oisín Mac Aodha, Simon Julier, and Gabriel J Brostow. “Structured prediction of unobserved voxels from a single depth image”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 5431–5440 (cit. on p. 1).
- [10] Jorge Fuentes-Pacheco, José Ruiz-Ascencio, and Juan Manuel Rendón-Mancha. “Visual simultaneous localization and mapping: a survey”. In: *Artificial Intelligence Review* 43 (2015) (cit. on p. 1).
- [11] Y. Furukawa, B. Curless, S.M. Seitz, and R. Szeliski. “Towards Internet-scale multi-view stereo”. In: *CVPR*. 2010, pp. 1434–1441 (cit. on p. 1).
- [12] Yasutaka Furukawa and Jean Ponce. “Accurate, Dense and Robust Multiview Stereopsis”. In: *PAMI* 32.8 (2010), pp. 1362–1376 (cit. on p. 1).
- [13] R. Girdhar, D.F. Fouhey, M. Rodriguez, and A. Gupta. “Learning a Predictable and Generative Vector Representation for Objects”. In: *ECCV*. 2016 (cit. on pp. 2, 3).
- [14] Michael Goesele, Jens Ackermann, Simon Fuhrmann, Ronny Klawnsky, Fabian Langguth, Patrick Müandcke, and Martin Ritz. “Scene Reconstruction from Community Photo Collections”. In: *IEEE Computer* 43 (6 2010), pp. 48–53 (cit. on p. 1).
- [15] JunYoung Gwak, Christopher B Choy, Manmohan Chandraker, Animesh Garg, and Silvio Savarese. “Weakly Supervised Generative Adversarial Networks for 3D Reconstruction”. In: *arXiv preprint* (2017) (cit. on pp. 1, 6).
- [16] Klaus Häming and Gabriele Peters. “The structure-from-motion reconstruction pipeline—a survey with focus on short image sequences”. In: *Kybernetika* 46.5 (2010), pp. 926–937 (cit. on p. 1).
- [17] Gleen Healey and Thomas O Binford. “Local shape from specularities”. In: *Computer Vision, Graphics, and Image Processing* 42.1 (1988), pp. 62–86 (cit. on p. 1).
- [18] Carlos Hernández and George Vogiatzis. “Shape from Photographs: A Multi-view Stereo Pipeline”. In: *Computer Vision*. Vol. 285. Studies in Computational Intelligence. Springer, 2010, pp. 281–311 (cit. on p. 1).
- [19] Qixing Huang, Hai Wang, and Vladlen Koltun. “Single-view reconstruction via joint analysis of image and shape collections”. In: *ACM Transactions on Graphics (TOG)* 34.4 (2015), p. 87 (cit. on p. 3).
- [20] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. “Spatial Transformer Networks”. In: *CoRR* abs/1506.02025 (2015) (cit. on pp. 2, 3).
- [21] Angjoo Kanazawa, David W. Jacobs, and Manmohan Chandraker. “WarpNet: Weakly Supervised Matching for Single-View Reconstruction”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 2016 (cit. on p. 3).
- [22] Maxime Lhuillier and Long Quan. “A quasi-dense approach to surface reconstruction from uncalibrated images”. In: *IEEE transactions on pattern analysis and machine intelligence* 27.3 (2005), pp. 418–433 (cit. on p. 1).
- [23] Jitendra Malik and Ruth Rosenholtz. “Computing local surface orientation and shape from texture for curved surfaces”. In: *International journal of computer vision* 23.2 (1997), pp. 149–168 (cit. on p. 1).

- [24] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. “Deep metric learning via lifted structured feature embedding”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 4004–4012 (cit. on p. 4).
- [25] Danilo Jimenez Rezende, SM Ali Eslami, Shakir Mohamed, Peter Battaglia, Max Jaderberg, and Nicolas Heess. “Unsupervised learning of 3d structure from images”. In: *Advances In Neural Information Processing Systems*. 2016, pp. 4997–5005 (cit. on pp. 1, 3).
- [26] Jason Rock, Tanmay Gupta, Justin Thorsen, JunYoung Gwak, Daeyun Shin, and Derek Hoiem. “Completing 3D object shape from one depth image”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 2484–2493 (cit. on p. 1).
- [27] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. “The earth mover’s distance as a metric for image retrieval”. In: *International journal of computer vision* 40.2 (2000), pp. 99–121 (cit. on p. 5).
- [28] Silvio Savarese, Marco Andreetto, Holly Rushmeier, Fausto Bernardini, and Pietro Perona. “3D reconstruction by shadow carving: Theory and practical evaluation”. In: *International journal of computer vision* 71.3 (2007), pp. 305–336 (cit. on p. 1).
- [29] Florian Schroff, Dmitry Kalenichenko, and James Philbin. “Facenet: A unified embedding for face recognition and clustering”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 815–823 (cit. on p. 4).
- [30] Thomas W Sederberg and Scott R Parry. “Free-form deformation of solid geometric models”. In: *ACM SIGGRAPH computer graphics* 20.4 (1986), pp. 151–160 (cit. on p. 5).
- [31] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. “Going deeper with convolutions”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 1–9 (cit. on p. 5).
- [32] Shubham Tulsiani, Tinghui Zhou, Alexei A. Efros, and Jitendra Malik. “Multi-view Supervision for Single-view Reconstruction via Differentiable Ray Consistency”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2017 (cit. on p. 1).
- [33] Jiajun Wu, Tianfan Xue, Joseph J Lim, Yuandong Tian, Joshua B Tenenbaum, Antonio Torralba, and William T Freeman. “Single image 3d interpreter network”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 365–382 (cit. on p. 1).
- [34] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T Freeman, and Joshua B Tenenbaum. “Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling”. In: *Neural Information Processing Systems (NIPS)*. 2016 (cit. on p. 3).
- [35] 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: *European Conference Computer Vision (ECCV)*. 2016 (cit. on p. 1).
- [36] 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*. IEEE. 2014 (cit. on p. 1).
- [37] 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*. 2016, pp. 1696–1704 (cit. on pp. 1, 3).
- [38] M. E. Yumer and N. J. Mitra. “Learning Semantic Deformation Flows with 3D Convolutional Networks”. In: *European Conference on Computer Vision (ECCV 2016)*. Springer. 2016 (cit. on pp. 3–5).
- [39] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. “Shape-from-shading: a survey”. In: *IEEE transactions on pattern analysis and machine intelligence* 21.8 (1999), pp. 690–706 (cit. on p. 1).

6 Supplementary Material

6.1 CD vs EMD

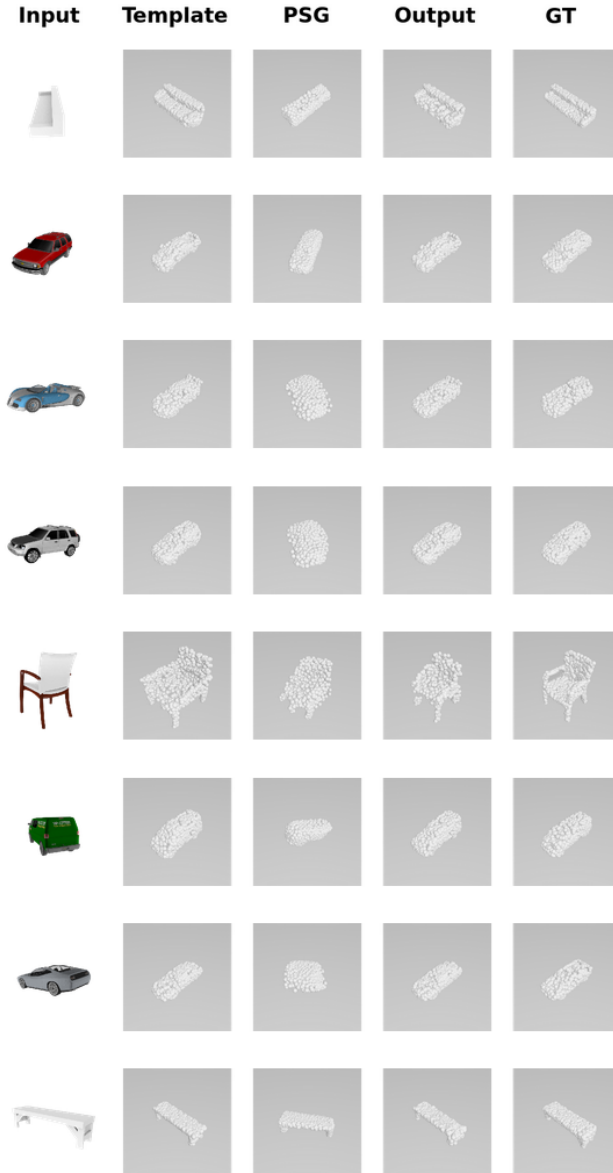
We trained DeformNet trained with both CD and EMD as objective functions, then test both of them using CD and EMD metrics. The average distance is reported on the Table 3. In terms of the point-set based metrics, the performance of these two models are similar. Besides, from our observation on output visualization, we don't find significant difference between these two metrics. With respect to computation efficiency, CD runs much faster with KDTree searching for nearest neighbors, so we primarily used it for the evaluations in this paper.

Table 3: Comparison of performance with CD and EMD. This evaluation is on a slightly different test set than in Table 1 and 2.

Train \ Test	CD	EMD
CD	0.10	0.52
EMD	0.11	0.50

6.2 Additional Images

We present 18 additional qualitative examples, drawn at random from the test set:



Input	Template	PSG	Output	GT
