3D Shape Completion with Multi-view Consistent Inference

Tao Hu, Zhizhong Han, Matthias Zwicker
Department of Computer Science
University of Maryland, College Park
taohu@cs.umd.edu, h312h@umd.edu, zwicker@cs.umd.edu

Abstract

3D shape completion is important to enable machines to perceive the complete geometry of objects from partial observations. To address this problem, view-based methods have been presented. These methods represent shapes as multiple depth images, which can be back-projected to yield corresponding 3D point clouds, and they perform shape completion by learning to complete each depth image using neural networks. While view-based methods lead to state-of-the-art results, they currently do not enforce geometric consistency among the completed views during the inference stage. To resolve this issue, we propose a multi-view consistent inference technique for 3D shape completion, which we express as an energy minimization problem including a data term and a regularization term. We formulate the regularization term as a consistency loss that encourages geometric consistency among multiple views, while the data term guarantees that the optimized views do not drift away too much from a learned shape descriptor. Experimental results demonstrate that our method completes shapes more accurately than previous techniques.

Introduction

Convolutional neural networks have proven highly successful at analysis and synthesis of visual data such as images and videos. This has spurred interest in applying convolutional network architectures also to 3D shapes, where a key challenge is to find suitable generalizations of discrete convolutions to the 3D domain. Popular techniques include using discrete convolutions on 3D grids (Wu et al. 2015), graph convolutions on meshes (Litany et al. 2018), convolution-like operators on 3D point clouds (Atzmon, Maron, and Lipman 2018; Li et al. 2018b), or 2D convolutions on 2D shape parameterizations (Cohen et al. 2018). A simple approach in the last category is to represent shapes using multiple 2D projections, or multiple depth images, and apply 2D convolutions on these views. This has led to successful techniques for shape classification (Su et al. 2015), single-view 3D reconstruction (Richter and Roth 2018), shape completion (Hu et al. 2019), and shape synthesis (Soltani et al. 2017). One issue in these approaches, however, is to encourage consistency among the separate views and avoid that each view represents a slightly different object. This is not an issue in supervised training, where the loss encourages all views to match the ground truth shape. But at inference time or in unsupervised training, ground truth is not available and a different mechanism is required to encourage consistency.

In this paper, we address the problem of shape completion using a multi-view depth image representation, and we propose a multi-view consistency loss that is minimized during inference. We formulate inference as an energy minimization problem, where the energy is the sum of a data term given by a conditional generative net, and a regularization term given by a geometric consistency loss. Our results show the benefits of optimizing geometric consistency in a multi-view shape representation during inference, and we demonstrate that our approach leads to state-of-the-art results in shape completion benchmarks. In summary, our contributions are as follows:

i) We propose a multi-view consistency loss for 3D shape completion that does not rely on ground truth data.
ii) We formulate multi-view consistent inference as an energy minimization problem including our consistency loss as a regularizer, and a neural network-based data term.
iii) We show state-of-the-art results in standard shape completion benchmarks, demonstrating the benefits of the multi-view consistency loss in practice.

Related Work

3D shape completion. Different 3D shape representations have been applied in 3D shape completion, such as voxels, point clouds, and multiple views. Voxel-based representations are widely used in shape completion with 3D CNN, such as 3D-Encoder-Predictor CNNs (Dai, Qi, and Nießner 2017) and encoder-decoder CNN for patch-level geometry refinement (Han et al. 2017). However, computational complexity grows cubically as the voxel resolution increases, which severely limits the completion accuracy. To address this problem, several point cloud-based shape completion methods (Achlioptas et al. 2018; Yang et al. 2017; Yuan et al. 2018) have been proposed. The point completion network (PCN) (Yuan et al. 2018) is a current state-of-the-art approach that extends the PointNet architecture (Qi et al. 2017) to provide an encoder, followed by a multi-stage decoder that uses both fully connected (Achlioptas et al. 2018)
Figure 1: Overview of the multi-view consistent inference for 3D shape completion. Given a partial point cloud as input, we first render multiple incomplete views $X$, which form our shape representation of the incomplete input. To perform inference, we apply a conditional generative network $G$ to generate completed depth images $V$ based on a shape descriptor $z$ conditioned on $X$. As a key idea, we design our consistency loss $C$ to evaluate the geometric consistency among $V$. Intuitively, for all pixels in all views $V_i$ we find the distance to their approximate closest neighbor in the other views $V_j$, and sum up these distances to form $C$. Specifically, for each target view (e.g. $V_2$ in the figure) we reproject all completed depth images $V_\ell$ according to the pose of $V_2$, which leads to reprojection maps denoted $R_{\ell}^{V_2}$. Then we compute consistency distances, denoted $D_{\ell}^{V_2}$, for each reprojection map $R_{\ell}^{V_2}$ and the target $V_2$ via a pixel-wise closest point pooling operation. Finally, a consistency pooling operator aggregates all consistency distances $D_{\ell}^{V_2}$ into a loss map $M^{V_2}$. In inference, we minimize all loss maps as a function of the shape descriptor $z$.

and folding layers (Yang et al. 2017). The output point cloud size in these methods is fixed, however, to small numbers like 2048 (Yang et al. 2017), which often leads to the loss of detail. View-based methods resolve this issue by completing each rendered depth image (Hu et al. 2019) of the incomplete shape, and then back-projecting the completed images into a dense point cloud. By leveraging state-of-the-art image-to-image translation networks (Isola et al. 2017), MVCN (Hu et al. 2019) completed each single view with a shape descriptor which encodes the characteristics of the whole 3D object to achieve higher accuracy. However, view-based methods fail to maintain geometric consistency among completed views during inference. Our approach resolves this issue using our novel multi-view consistent inference technique.

**Multi-view consistency.** One problem of view-based representation is inconsistency among multiple views. Some researchers presented a multi-view approach to train their network to achieve consistency in multi-view representations, like discovering 3D keypoints (Suwajanakorn et al. 2018) and reconstructing 3D objects from images (Lin, Kong, and Lucey 2018; Li et al. 2018a; Tulsiani, Efros, and Malik 2018; Jiang et al. 2018; Khot et al. 2019). With differentiable rendering (Lin, Kong, and Lucey 2018; Tulsiani, Efros, and Malik 2018), the consistency distances among different views can be leveraged as 2D supervision to learn 3D shapes in their networks. However, these methods can only guarantee consistency for training data in training stage. Different from these methods, with the help of our novel energy optimization and consistency loss implementation, our proposed method can improve geometric consistency on test data directly during the inference stage.

**Multi-view Consistent Inference**

**Overview.** The goal of our method is to guarantee multi-view consistency in inference, as shown in the overview in Fig. 1. Our method starts from converting partial point clouds to multi-view depth image representations by rendering the points into a set of depth images $X = \{X_1, \ldots, X_8\}$ from a number of fixed viewpoints. In our current implementation, we use eight viewpoints placed on the corners of a cube. Our approach builds on a conditional generative net $G(z; X)$ which is trained to output completed depth images $V$ by estimating a shape descriptor $z$ conditioned on a set of incomplete inputs $X$. We obtain the conditional generative net in a separate, supervised training stage. During inference, we keep the network weights fixed and optimize the shape descriptor $z$ to minimize an energy consisting of a consistency loss, which acts as a regularizer, and a data term. On the other hand, the consistency loss $C(V) = C(G(z; X))$ quantifies the geometric consistency among the completed depth images $V$. On the other hand, the data term encourages the solution to stay close to an initially estimated shape descriptor $\hat{z}$. This leads to the following optimization for the desired shape descriptor $z^*$:

$$z^* = \arg \min_z C(G(z; X)) + \mu \|G(z; X) - G(\hat{z}; X)\|$$

$$= L_{con}(z) + \mu L_{gen}(z),$$

where $\mu$ is a weighting factor, and we denote $Y = G(\hat{z}; X)$ and $V = G(z; X)$ as the initially estimated completed depth.
images and optimized completed depth images in inference, respectively. In addition, we will formalize the regularization term and data term as multi-view consistency loss $\mathcal{L}_{con}(z)$ and generator loss $\mathcal{L}_{gen}(z)$ in Section ‘Consistency Loss’.

**Conditional generative net.** The conditional generative net $G(z; X)$ is built on the structure of multi-view completion net (Hu et al. 2019), as shown in Fig. 2, which is an image-to-image translation architecture applied to perform depth image completion for multiple views of the same shape. We train the conditional generative net following a standard conditional GAN approach (Goodfellow et al. 2014). To share information between multiple depth images of the same shape, our architecture learns a shape descriptor $z$ for each 3D object by pooling a so-called shape memory consisting of $N$ feature maps $f_n$, $n \in [1, N = 8]$ from all views of the shape. The network $G$ consists of 8 U-Net modules, and each U-Net module has two submodules, Down and Up, so there are 8 Down submodules ($D_{7-0}$) in the encoder and 8 Up submodules ($U_{0-7}$) in the decoder. Down submodules consist of the form Convolution-BatchNorm-ReLU (Ioffe and Szegedy 2015; Nair and Hinton 2010), and Up submodules ($D_{0-7}$) consist of the form UpReLU-UpConv-UpNorm. The shape memory is the feature map after the third Down submodule ($D_3$) of the encoder. More details can be found in (Isola et al. 2017; Hu et al. 2019).

In inference, we optimize the shape descriptor $z$ of $G(z; X)$ given test input $X$. We first get an initial estimation of the shape descriptor $\hat{z}$ for each test shape by running the trained model once, and initialize $z$ with $\hat{z}$. During inference the other parameters of $G$ are fixed.

**Consistency Loss**

Our consistency loss is based on the sum of the distances between each pixel in the multi-view depth map and its approximate nearest neighbor in any of the other views. In this section we introduce the details of the multi-view consistency loss calculation following the overview in Fig. 1. For all views $V_t$, we first calculate pairwise per-pixel consistency distances $D_s^t$ to each other view $V_s$, that is, per-pixel distances to approximate nearest neighbors in view $V_s$. We then perform consistency pooling, which for each view $V_t$ provides the consistency distances over all other views (as opposed to the initial pairwise consistency distances between two of views). We call these the loss maps $M^t$. The final consistency loss is the sum over all loss maps.

**Pairwise Consistency Distances**

Given a source view $V_s$ and a target view $V_t$, we calculate the consistency distance $D_s^t$ between $V_s$ and $V_t$ by view-reprojection and closest point pooling, where $V_t, V_s \in \mathbb{R}^{H \times W}$ and $H \times W$ is the image resolution. Specifically, view-reprojection transforms the depth information of source $V_s$ to a reprojection map $R_s^t$ according to the transformation matrix of the target $V_t$. Then, closest point pooling further produces the consistency distance $D_s^t$ between $R_s^t$ and $V_t$. Fig. 3 shows the pipeline, where the target view is $V_t$ and the source view is $V_s$. In the following, we denote a pixel on source view as $p_i = [u_i, v_i, d_i]$, where $u_i$ and $v_i$ are considered pixel coordinates, its back-projected 3D point as $P_i = [\hat{x}_i, \hat{y}_i, \hat{z}_i]$, and the reprojected pixel on reprojection map $R_s^t$ as $p_i^t = [u'_i, v'_i, d'_i]$, where $d_i = V_s[u_i, v_i]$ and $d'_i = R_s^t[u'_i, v'_i]$ are the depth values at the location $[u_i, v_i]$ and $[u'_i, v'_i]$, respectively.

**View-reprojection.** The view-reprojection operator back-projects each point $p_i = [u_i, v_i, d_i]$ on $V_s$ into the canonical 3D coordinates as $P_i = [\hat{x}_i, \hat{y}_i, \hat{z}_i]$ via

$$P_i = \mathbb{R}_s^{-1}(K^{-1}p_i - \tau_s) \quad \forall i,$$

where $K$ is the intrinsic camera matrix, and $\mathbb{R}_s$ and $\tau_s$ are the rotation matrix and translation vector of view $V_s$ respectively. This defines the relationship between the view $V_s = \{p_i\}$ and its back-projected point cloud $\{P_i\}$. We use $T_s$ to denote the transformation matrix of $V_s$, which contains the pose information, such that $T_s = (\mathbb{R}_s, \tau_s)$. Then, we transform each 3D point $P_i$ in the point cloud into a pixel $p_i^t = [u'_i, v'_i, d'_i]$ on the reprojection map $R_s^t$ as

$$p_i^t = K(\mathbb{R}_sP_i + \tau_s) \quad \forall i.$$

Eq. (2) and Eq. (3) illustrate that we can transform the depth information of source view $V_s$ to reprojection map $R_s^t$, which has the same pose with the target view $V_t$. However, due to the discrete grid of the depth images, different points $P_i$ in the point cloud may be projected to the same pixel $[u'_i, v'_i]$ on the reprojection map $R_s^t$ when using Eq. (3), like $p_1^t = [u'_1, v'_1, d'_1], p_2^t = [u'_1, v'_1, d'_2], p_3^t = [u'_1, v'_1, d'_3]$ in Fig. 3. In fact, all the $\{p_1^t, p_2^t, p_3^t\}$ are projected to the same pixel $p_1^t$ on $R_s^t$, and the corresponding point on the target view $V_t$ is $p_t$. To alleviate this collision effect, we implement a
pseudo-rendering technique similar to (Lin, Kong, and Lucey 2018). Specifically, for each pixel on \(R^t_s\), a sub-pixel grid with a size of \((U \times U)\) is presented to store multiple depth values corresponding to the same pixel, so the reprojection is \(R^t_s \in \mathbb{R}^{H \times U \times W \times U}\).

**Closest point pooling.** The closest point pooling operator computes the consistency distance between reprojection \(R^t_s\) and target view \(V_t\). First, we also upsample \(V_t\) to \(\mathbb{R}^{H \times U \times W \times U}\) by repeating each depth value into a \(U \times U\) sub-pixel grid. Then, we calculate the element-wise \(L_1\) distance between \(R^t_s\) and the upsampled \(V_t\). Finally, we perform closest point pooling to extract the minimal \(L_1\) distance in each sub-pixel grid using min-pooling with a \(U \times U\) filter and a stride of \(U \times U\). This provides the consistency distance \(D^t_s\) between source view \(V_s\) and target view \(V_t\), where \(D^t_s \in \mathbb{R}^{H \times W}\). The consistency distance \(D^t_s\) is shown in Fig. 3, where \(t = 7, s = 2\). Note that we directly take the \(t\)th input view \(V_t\) as the reprojection \(R^t_s\) when \(t = s\), since the incomplete input \(X_t\) also provides some supervision.

Note that some consistency distances in \(D^t_s\) may be large due to noisy view completion or self-occlusion between the source and target views, and these outliers interfere with our energy minimization. Therefore, we perform outlier suppression by ignoring consistency distances above a threshold of 2.5% of the depth range (from the minimum to the maximum depth value of a model).

**Consistency Distance Aggregation by Consistency Pooling**

Given a target view \(V_t\), we get all the consistency distances \(D^t_s\) between \(V_t\) and all the other \(N\) source views \(V_s\), as shown in Fig. 4, where \(t = 7, N = 8\), and we use the same colorbar with Fig. 1. Obviously, different source views \(V_s\) cover different parts of the target view \(V_t\), which leads to different consistency distances in \(D^t_s\). For example, the red parts on each \(D^t_s\) in Fig. 4 indicate that they cannot be well inferred from the source view, so these parts are not helpful for the optimization of the target view.

By extracting the minimum distance between \(V_t\) and the reprojections from all other views, we cover the whole \(V_t\) with the closest points to it and we obtain the loss maps \(M^t\). In our pipeline, we implement this efficiently using a consistency pooling operator defined as,

\[
M^t(x, y) = \min_{j \in [1, J]} D^t_j(x, y), \tag{4}
\]

where \(M^t \in \mathbb{R}^{H \times W}\), \(x \in [1, H], y \in [1, W]\), and \(J\) is the number of views in pooling. We use \(J \leq N\) to make it possible to restrict pooling to a subset of the views (see Section ‘Experiments’ for an evaluation of this parameter). This is illustrated using \(M^t\) as an example in Fig. 4. Fig. 5 shows all the consistency loss maps to each target view.

**Loss Function**

Our multi-view consistent inference aims to maximize the depth consistency across all views by optimizing the shape descriptor \(z\) of a 3D model. Therefore, the consistency loss \(L_{con}(z)\) to the whole 3D model takes the loss maps for all target views,

\[
L_{con}(z) = C(G(z; X)) = \frac{1}{N \times H \times W} \sum_{t=1}^{N} \sum_{x=1}^{H} \sum_{y=1}^{W} M^t(x, y), \tag{5}
\]

where \(N\) is the number of views and \(X\) is the input set of incomplete depth images of the 3D model.

In Eq. (1), we also have a data term to keep \(z\) close to the initial estimation \(\hat{z}\) during inference. We call this the generator loss \(L_{gen}(z)\), which aims to prevent the completed depth images drifting away from the prior learned from the training data:

\[
L_{gen}(z) = \|G(z; X) - G(\hat{z}; X)\|, \tag{6}
\]

where \(X\) is the input, \(Y = G(\hat{z}; X)\) and \(V = G(z; X)\) are the initially estimated outputs and optimized outputs respectively, and \(X, Y, V \in \mathbb{R}^{N \times H \times W}\). In summary, the overall loss function in inference \(L(z)\) is

\[
L(z) = L_{con}(z) + \mu L_{gen}(z), \tag{7}
\]

where \(\mu\) is a weighting factor. We optimize the shape descriptor \(z\) for 100 gradient descent steps, and we take \(z\) with the smallest consistency loss in the last 10 steps as \(z^*\). It should be mentioned that since the gradients of \(z\) are small, we use a large learning rate of 0.2.

**Experiments**

Our method is built on MVCN (Hu et al. 2019), a state-of-the-art view-based shape completion method. To fairly evaluate the improvements over MVCN directly, we use the same pipeline a MVCN to generate training and test depth images, where each 3D object is represented by \(N = 8\) depth maps with a resolution of \(256 \times 256\). We take 3D models from ShapeNet (Chang et al. 2015). Initially, we set \(J = N\) in Eq. 5 to conduct consistency pooling in the following experiments. In addition, we use the same training dataset and hyperparameters with (Hu et al. 2019) to train the network, and the same test dataset with (Hu et al. 2019; Yuan et al. 2018) to evaluate our methods with Chamfer Distance (CD) (Fan, Su, and Guibas 2017).

**Analysis of the Objective Function**

We test different objective functions in Eq. (7) to justify the effectiveness of our methods. Table 1 shows the quantitative effects of these variations. The experiments are conducted on 100 3D airplane models (besides test dataset or training dataset), which are randomly selected under the constraints that the average CD is close to that of the test dataset in (Hu et al. 2019). We change the weighting factor \(\mu\) between \(L_{con}(z)\) and \(L_{gen}(z)\), and different distance functions in \(L_{gen}(z)\) (using \(L_1\) or \(L_2\)). When \(\mu = 0\), only \(L_{con}(z)\) is used in loss function. According to the comparison, we select \(L_2\) distance to calculate generator loss, and set \(\mu = 1\) in the following experiments.

**The Size of Depth-buffer in Pseudo-rendering**

As mentioned above, we use a depth-buffer in pseudo-rendering, and the depth-buffer size is \(U \times U\). Obviously,
Table 1: Chamfer distance over different loss functions in Eq. 7. CD is multiplied by $10^3$.

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$\mu = 0.1$</th>
<th>$\mu = 1$</th>
<th>$\mu = 2$</th>
<th>$\mu = 5$</th>
<th>$\mu = 10$</th>
<th>$\mu = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_1$</td>
<td>5.228</td>
<td>5.160</td>
<td>5.129</td>
<td>5.160</td>
<td>5.155</td>
<td>6.383</td>
</tr>
<tr>
<td>$L_2$</td>
<td>5.362</td>
<td>5.110</td>
<td>5.136</td>
<td>5.135</td>
<td>5.175</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: The effects of depth-buffer sizes $U$ (a) and numbers of views $J$ (in Eq. 4) in consistency pooling (b). CD is multiplied by 100.

<table>
<thead>
<tr>
<th>$U$</th>
<th>Average CD</th>
<th>$J$</th>
<th>Average CD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Table</td>
<td>Sofa</td>
<td></td>
</tr>
<tr>
<td>$U = 1$</td>
<td>0.8876</td>
<td>0.8440</td>
<td></td>
</tr>
<tr>
<td>$U = 3$</td>
<td>0.8830</td>
<td>0.8421</td>
<td></td>
</tr>
<tr>
<td>$U = 5$</td>
<td>0.8754</td>
<td>0.8394</td>
<td></td>
</tr>
</tbody>
</table>

(a) (b)

The Number of Views in Consistency Pooling

In this part, we analyze the effects of varying the number of views $J$ in consistency pooling. As shown in Fig. 4, more views mean a bigger coverage over the target view and a smaller consistency loss. Given a depth-buffer size of $5 \times 5$, Fig. 6 (e) to (g) show that the consistency loss increases when $J = 3$ or $J = 5$, and we also find more noisy points in these loss maps Fig. 6 (f) and (g).

Comparison with Direct Optimization Method

Our multi-view consistent inference can also be used to optimize completed depth maps directly without the conditional generative net $G$. We call this direct optimization on depth maps, and in this part, we compare our methods with direct optimization. In fact, direct optimization only contains the Consistency loss calculation $C$ part in Fig. 1. Each depth map will be a trainable tensor. We first initialize the tensors with the completed views $V_n, n \in [1,8]$, and then update these tensors by minimizing the consistency loss in Eq. 7. We use $L_2$ distance to calculate $L_{gen}(z), \mu = 1$, and the learning rate is 0.0006, which produces the best results for direct optimization. Fig. 7 shows the comparisons. Here we color-code the normals of the complete point clouds, which are estimated using a k-d tree search algorithm with a search radius of 0.5 and a maximum number of neighbors of 30. Compared with direct optimization, our method performs better. For example, in terms of optimizing point clouds, we can smooth the surface, like the seat of the chair, and remove some outliers. As for completing depth maps, our method can fill a hole appearing in MVCN (Hu et al. 2019) and even add the missing leg, where the $L_1$ distance to the ground truth are marked in red.

Though the direct optimization method can also refine the point clouds of MVCN, it does not perform well in removing outliers on point clouds (left) or completing a depth map (right) in Fig. 7. The reason is that direct optimization does not have any knowledge to distinguish shape and background from a depth map, which means that for pixels in a hole, direct optimization does not know whether they belong to a hole of the shape or the background. However, with the knowledge of shape completion learned in the conditional generative net $G$, our method completes shapes better.

Intermediate Results and Convergence

In Fig. 8, the image insets illustrate the intermediate completion results of the [0, 20, 40, 60, 80, 100]th step for one example depth image from the cabinet class. In addition, $L_{gt}(z) = \|GT - G(z; X)\|$ is averaged over all cabinet objects, where $GT$ is ground truth. For clarity, the curve is offset vertically by 0.2. $\Delta L_{gt}(z) = \|GT - G(z; X)\| - \|GT - G(z; X)\|$. We see the completed results are closer to ground truth than MVCN, though there is no ground truth supervision in inference.

Fig. 8 illustrates empirically that, under the defined loss function, our optimization can find a good solution within 100 steps. The figure shows the average loss vs gradient descent steps on all the 150 cabinet test objects. We reach the maximum of $L_{gen}(z)$ within $s$ steps, then the distance to $G(z; X)$ decreases in the following $100 - s$ steps. For 98% of all the 1200 test objects, the maximum is reached within 10 steps ($s < 10$), and within 20 steps for almost all. After 100 steps, the optimization has largely converged.
Figure 6: Consistency loss maps over different depth-buffer sizes \((U)\) and numbers of views in consistency pooling \(J\). (c) to (g) are the consistency loss maps, where the values of the consistency loss (scaled with 100) are marked in red. We use the same colorbar with Fig. 1.

Figure 7: Comparisons between direct optimization and our methods on optimizing point clouds (left) and depth maps (right). The normals of the point clouds are shown.

Figure 8: Consistent inference optimization (loss vs steps).

**Completion results**

**Improvements over Existing Works.** Here we compare our method with the state-of-the-art shape completion methods, including 3D-EPN (Dai, Qi, and Nießner 2017), FC (Achlioptas et al. 2018), Folding (Yang et al. 2017), three variants of PCN (Yuan et al. 2018): PN2, PCN-CD, PCN-EMD, and MVCN (Hu et al. 2019). TopNet (Tchapmi et al. 2019) is a recent point-based method, but their generated point clouds are sparse.

Table 3 shows the quantitative results, where the completion results of the other methods are from (Yuan et al. 2018; Hu et al. 2019) and ‘Direct-Opt’ is the direct optimization method introduced above. With multi-view consistency optimization, both direct optimization and our method can improve MVCN on most categories of the test datasets, and our method achieves better results. The optimization methods fail on the Lamp dataset. As mentioned in (Hu et al. 2019), the reason is that the completion of MVCN is bad on several lamp objects, which makes the optimization less meaningful.

Fig. 11 shows the qualitative improvements over the currently best view-based method, MVCN, where the normals of point clouds are color-coded. With the conditional generative net \(G\) and multi-view consistency loss \(C\), our method produces completed point clouds with smoother surfaces and fewer outliers, and can also fill holes of shapes on multiple categories.

**Completions results given different inputs.** Fig. 10 (a, b, c) show completed airplanes and cars under 3 different inputs of the same objects. Since the car input in (a) leaves a lot of ambiguity, the completed cars vary. The airplanes results are more similar because the inputs contain most of the structure.

**Multiple views of completed shapes.** Fig 10 (c) shows a completed airplane and car from 3 views. We see the completed shapes are consistent among different views.

Figure 9: Completions on noisy inputs. GT is ground truth.

**Completions on noisy inputs.** In Fig. 9, we perturb the input depth map with Gaussian noise whose standard deviation is 0.01 times the scale of the depth measurements. Our completion is robust to the noisy input.

**Conclusion**

We proposed multi-view consistent inference to enforce geometric consistency in view-based 3D shape completion. We defined a novel multi-view consistency loss suitable for optimization in inference, which can be achieved without the supervision of ground truth. The experimental results demonstrate that our method can complete 3D shapes more accurately than existing methods.
Table 3: Mean Chamfer Distance over multiple categories in ShapeNet. CD is scaled by 100.

<table>
<thead>
<tr>
<th>Method</th>
<th>Avg</th>
<th>Airplane</th>
<th>Cabinet</th>
<th>Car</th>
<th>Chair</th>
<th>Lamp</th>
<th>Sofa</th>
<th>Table</th>
<th>Vessel</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D-EPN</td>
<td>2.0147</td>
<td>1.3161</td>
<td>2.1803</td>
<td>2.0306</td>
<td>1.8813</td>
<td>2.5746</td>
<td>2.1089</td>
<td>2.1716</td>
<td>1.8543</td>
</tr>
<tr>
<td>FC</td>
<td>0.9799</td>
<td>0.5698</td>
<td>1.1023</td>
<td>0.8775</td>
<td>1.0969</td>
<td>1.1131</td>
<td>1.1756</td>
<td>0.9320</td>
<td>0.9720</td>
</tr>
<tr>
<td>Folding</td>
<td>1.0074</td>
<td>0.5965</td>
<td>1.0831</td>
<td>0.9272</td>
<td>1.1245</td>
<td>1.2172</td>
<td>1.1630</td>
<td>0.9453</td>
<td>1.0027</td>
</tr>
<tr>
<td>PN2</td>
<td>1.3999</td>
<td>1.0300</td>
<td>1.4735</td>
<td>1.2187</td>
<td>1.5775</td>
<td>1.7615</td>
<td>1.6183</td>
<td>1.1676</td>
<td>1.3521</td>
</tr>
<tr>
<td>PCN-CD</td>
<td>0.9636</td>
<td>0.5502</td>
<td>1.0625</td>
<td>0.8696</td>
<td>1.0998</td>
<td>1.1339</td>
<td>1.1676</td>
<td>0.8590</td>
<td>0.9665</td>
</tr>
<tr>
<td>PCN-EMD</td>
<td>1.0021</td>
<td>0.5849</td>
<td>1.0685</td>
<td>0.9080</td>
<td>1.1580</td>
<td>1.1961</td>
<td>1.2206</td>
<td>0.9014</td>
<td>0.9789</td>
</tr>
<tr>
<td>MVCN</td>
<td>0.8298</td>
<td>0.5273</td>
<td>0.7154</td>
<td>0.6322</td>
<td>1.0077</td>
<td>1.0576</td>
<td>0.9174</td>
<td>0.9020</td>
<td>0.8790</td>
</tr>
<tr>
<td>Direct-Opt</td>
<td>0.8195</td>
<td>0.5182</td>
<td>0.7001</td>
<td>0.6156</td>
<td>0.9820</td>
<td>1.1032</td>
<td>0.8885</td>
<td>0.8854</td>
<td>0.8619</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>0.8052</strong></td>
<td><strong>0.5175</strong></td>
<td><strong>0.6722</strong></td>
<td><strong>0.5817</strong></td>
<td><strong>0.9547</strong></td>
<td><strong>1.1334</strong></td>
<td><strong>0.8394</strong></td>
<td><strong>0.8754</strong></td>
<td><strong>0.8669</strong></td>
</tr>
</tbody>
</table>

Figure 10: Completion results given three different inputs (a, b, c). 3 different views (c). GT indicates ground truth (d).

Figure 11: Improvements over MVCN on multiple categories in ShapeNet. GT is ground truth.
References

Achlioptas, P.; Diamanti, O.; Mitliagkas, I.; and Guibas, L. J. 2018. Learning representations and generative models for 3d point clouds. In \textit{ICML}.


Litany, O.; Bronstein, A.; Bronstein, M.; and Makadia, A. 2018. Deformable shape completion with graph convolutional autoencoders. \textit{CVPR}.


Tchapmi, L. P.; Kosaraju, V.; Rezatofighi, H.; Reid, I.; and Savarese, S. 2019. Topnet: Structural point cloud decoder. In \textit{The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)}.


