DIST: Rendering Deep Implicit Signed Distance Function with Differentiable Sphere Tracing

Shaohui Liu\textsuperscript{1,3}† Yinda Zhang\textsuperscript{2} Songyou Peng\textsuperscript{1,6} Boxin Shi\textsuperscript{4,7} Marc Pollefeys\textsuperscript{3,5,6} Zhaopeng Cui\textsuperscript{1}∗

\textsuperscript{1}ETH Zurich \textsuperscript{2}Google \textsuperscript{3}Tsinghua University \textsuperscript{4}Peking University \textsuperscript{5}Microsoft \textsuperscript{6}Max Planck ETH Center for Learning Systems \textsuperscript{7}Peng Cheng Laboratory

Abstract

We propose a differentiable sphere tracing algorithm to bridge the gap between inverse graphics methods and the recently proposed deep learning based implicit signed distance function. Due to the nature of the implicit function, the rendering process requires tremendous function queries, which is particularly problematic when the function is represented as a neural network. We optimize both the forward and backward passes of our rendering layer to make it run efficiently with affordable memory consumption on a commodity graphics card. Our rendering method is fully differentiable such that losses can be directly computed on the rendered 2D observations, and the gradients can be propagated backwards to optimize the 3D geometry. We show that our rendering method can effectively reconstruct accurate 3D shapes from various inputs, such as sparse depth and multi-view images, through inverse optimization. With the geometry based reasoning, our 3D shape prediction methods show excellent generalization capability and robustness against various noises.

1. Introduction

Solving vision problem as an inverse graphics process is one of the most fundamental approaches, where the solution is the visual structure that best explains the given observations. In the realm of 3D geometry understanding, this approach has been used since the very early age \cite{1,35,54}. As a critical component to the inverse graphics based 3D geometric reasoning process, an efficient renderer is required to accurately simulate the observations, \textit{e.g.}, depth maps, from an optimizable 3D structure, and also be differentiable to back-propagate the error from the partial observation.

As a natural fit to the deep learning framework, differentiable rendering techniques have drawn great interests recently. Various solutions for different 3D representations, \textit{e.g.}, voxels, point clouds, meshes, have been proposed. However, these 3D representations are all discretized up to a certain resolution, leading to the loss of geometric details and breaking the differentiable properties \cite{23}. Recently, the continuous implicit function has been used to represent the signed distance field \cite{34}, which has premium capacity to encode accurate geometry when combined with the deep learning techniques. Given a latent code as the shape representation, the function can produce a signed distance value for any arbitrary point, and thus enable unlimited resolution and better preserved geometric details for rendering purpose. However, a differentiable rendering solution for learning-based continuous signed distance function does not exist yet.

In this paper, we propose a differentiable renderer for continuous implicit signed distance functions (SDF) to facilitate the 3D shape understanding via geometric reasoning in a deep learning framework (Fig. 1). Our method can render an implicit SDF represented by a neural network from a latent code into various 2D observations, \textit{e.g.}, depth images, surface normals, silhouettes, and other properties encoded, from arbitrary camera viewpoints. The rendering process is fully differentiable, such that loss functions
can be conveniently defined on the rendered images and the observations, and the gradients can be propagated back to the shape generator. As major applications, our differentiable renderer can be applied to infer the 3D shape from various inputs, e.g., multi-view images and single depth image, through an inverse graphics process. Specifically, given a pre-trained generative model, e.g., DeepSDF [34], we search within the latent code space for the 3D shape that produces the rendered images mostly consistent with the observation. Extensive experiments show that our geometric reasoning based approach exhibits significantly better generalization capability than previous purely learning based approaches, and consistently produce accurate 3D shapes across datasets without finetuning.

Nevertheless, it is challenging to make differentiable rendering work on a learning-based implicit SDF with computationally affordable resources. The main obstacle is that an implicit function provides neither the exact location nor any bound of the surface geometry as in other representations like meshes, voxels, and point clouds.

Inspired by traditional ray-tracing based approaches, we adopt the sphere tracing algorithm [13], which marches along each pixel’s ray direction with the queried signed distance until the ray hits the surface, i.e., the signed distance equals to zero (Fig. 2). However, this is not feasible in the neural network based scenario where each query on the ray would require a forward pass and recursive computational graph for back-propagation, which is prohibitive in terms of computation and memory.

To make it work efficiently on a commodity level GPU, we optimize the full life-time of the rendering process for both forward and backward propagations. In the forward pass, i.e., the rendering process, we adopt a coarse-to-fine approach to save computation at initial steps, an aggressive strategy to speed up the ray marching, and a safe convergence criteria to prevent unnecessary queries and maintain resolution. In the backward propagation, we propose a gradient approximation which empirically has negligible impact on the training performance but dramatically reduces the computation and memory consumption. By making the rendering tractable, we show how producing 2D observations with the sphere tracing and interacting with camera extrinsics can be done in differentiable ways.

To sum up, our major contribution is to enable efficient differentiable rendering on the implicit signed distance function represented as a neural network. It enables accurate 3D shape prediction via geometric reasoning in deep learning frameworks and exhibits promising generalization capability. The differentiable renderer could also potentially benefit various vision problems thanks to the marriage of implicit SDF and inverse graphics techniques. The code and data are available at https://github.com/B1ueber2y/DIST-Renderer.

2. Related Work

3D Representation for Shape Learning. The study of 3D representations for shape learning is one of the main focuses in 3D deep learning community. Early work quantizes shapes into 3D voxels, where each voxel contains either a binary occupancy status (occupied / not occupied) [52, 6, 46, 39, 12] or a signed distance value [55, 9, 45]. While voxels are the most straightforward extension from the 2D image domain into the 3D geometry domain for neural network operations, they normally require huge memory overhead and result in relatively low resolutions. meshes are also proposed as a more memory efficient representation for 3D shape learning [47, 11, 22, 20], but the topology of meshes is normally fixed and simple. Many deep learning methods also utilize point clouds as the 3D representation [37, 38]; however, the point-based representation lacks the topology information and thus makes it non-trivial to generate 3D meshes. Very recently, the implicit functions, e.g., continuous SDF and occupancy function, are exploited as 3D representations and show much promising performance in terms of the high-frequency detail modeling and the high resolution [34, 29, 30, 4]. Similar idea has also been used to encode other information such as textures [33, 40] and 4D dynamics [32]. Our work aims to design an efficient and differentiable renderer for the implicit SDF-based representation.

Differentiable Rendering. With the success of deep learning, the differentiable rendering starts to draw more attention as it is essential for end-to-end training, and solutions have been proposed for various 3D representations. Early works focus on 3D triangulated meshes and leverage standard rasterization [28]. Various approaches try to solve the discontinuity issue near triangle boundaries by smoothing the loss function or approximating the gradient [21, 36, 25, 3]. Solutions for point clouds and 3D voxels are also introduced [48, 17, 31] to work jointly with PointNet [37] and 3D convolutional architectures. However, the differentiable rendering for the implicit continuous function representation does not exist yet. Some ray tracing based approaches are related, while they are mostly proposed for explicit representation, such as 3D voxels [27, 31, 43, 18] or meshes [23], but not implicit functions. Liu et al. [26] firstly propose to learn from 2D observations over occupancy networks [29]. However, their methods make several approximations and do not benefit from the efficiency of rendering implicit SDF. Most related to our work, Sitzmann et al. [44] propose a LSTM-based renderer for an implicit scene representation to generate color images, while their model focuses on simulating the rendering process with an LSTM without clear geometric meaning. This method can only generate low-resolution images due to the expensive memory consumption. Alternatively, our method can directly render 3D geometry represented by an implicit SDF.
to produce high-resolution images. It can also be applied without training to existing deep learning models.

**3D Shape Prediction.** 3D shape prediction from 2D observations is one of the fundamental vision problems. Early works mainly focus on multi-view reconstruction using multi-view stereo methods [41, 14, 42]. These purely geometry-based methods suffer from degraded performance on texture-less regions without prior knowledge [7]. With progress of deep learning, 3D shapes can be recovered under different settings. The simplest setting is to recover 3D shape from a single image [6, 10, 51, 19]. These systems rely heavily on priors, and are prone to weak generalization. Deep learning based multi-view shape prediction methods [53, 15, 16, 49, 50] further involve geometric constraints across views in the deep learning framework, which shows better generalization. Another thread of works [9, 8] take a single depth image as input, and the problem is usually referred as shape completion. Given the shape prior encoded in the neural network [34], our rendering method can effectively predict accurate 3D object shape from a random initial shape code with various inputs, such as depth and multi-view images, through geometric optimization.

### 3. Differentiable Sphere Tracing

In this section, we introduce our differentiable rendering method for the implicit signed distance function represented as a neural network, such as DeepSDF [34]. In DeepSDF, a network takes a latent code and a 3D location as input, and produces the corresponding signed distance value. Even though such a network can deliver high quality geometry, the explicit surface cannot be directly obtained and requires dense sampling in the 3D space.

Our method is inspired by Sphere Tracing [13] designed for rendering SDF volumes, where rays are shot from the camera pinhole along the direction of each pixel to search for the surface level set according to the signed distance value. However, it is prohibitive to apply this method directly on the implicit signed distance function represented as a neural network, since each tracing step needs a feed-forward neural network and the whole algorithm requires unaffordable computational and memory resources. To make this idea work in deep learning framework for inverse graphics, we optimize both the forward and backward propagations for efficient training and test-time optimization. The sphere traced results, i.e., the distance along the ray, can be converted into many desired outputs, e.g., depth, surface normal, silhouette, and hence losses can be conveniently applied in an end-to-end manner.

#### 3.1. Preliminaries - Sphere Tracing

To be self-contained, we first briefly introduce the traditional sphere tracing algorithm [13]. Sphere tracing is a conventional method specifically designed to render depth from volumetric signed distance fields. For each pixel on the image plane, as shown in Figure 2, a ray (L) is shot from the camera center (c) and marches along the direction (v̂) with a step size that is equal to the queried signed distance value (b). The ray marches iteratively until it hits or gets sufficiently close to the surface (i.e. abs(SDF) < threshold). A more detailed algorithm can be found in Algorithm 1.

#### 3.2. Efficient Forward Propagation

Directly applying sphere tracing to an implicit SDF function represented by a neural network is prohibitively computational expensive, because each query of f requires a forward pass of a neural network with considerable capacity. Naive parallelization is not sufficient since essentially millions of network queries are required for a single rendering with VGA resolution (640 × 480). Therefore, we need to cut off unnecessary marching steps and safely speed up the marching process.

**Initialization.** Because all the 3D shapes represented by DeepSDF are bounded within the unit sphere, we initialize p(0) to be the intersection between the camera ray and the unit sphere for each pixel. Pixels with the camera rays that do not intersect with the unit sphere are set as background (i.e., infinite depth).

**Coarse-to-fine Strategy.** At the beginning of sphere tracing, rays for different pixels are fairly close to each other, which indicates that they will likely march in a similar way. To leverage this nice property, we propose a coarse-to-fine

---

**Figure 2.** Illustration on the sphere tracing algorithm [13]. A ray is initiated at each pixel and marching along the viewing direction. The front end moves with a step size equals to the signed distance value of the current location. The algorithm converges when the current absolute SDF is smaller than a threshold, which indicates that the surface has been found.

---

**Algorithm 1** Naive sphere tracing algorithm for a camera ray L : c + d̂v over a signed distance fields f : N^3 → R.

1: Initialize n = 0, d(n) = 0, p(n) = c.
2: while not converged do:
3: Take the corresponding SDF value b(n) = f(p(n)) of the location p(n) and make update: d(n+1) = d(n) + b(n).
4: p(n+1) = c + d(n+1)v, n = n + 1.
5: Check convergence.
6: end while
sphere tracing strategy as shown in Fig. 3 (a). We start the sphere tracing from an image with \( \frac{1}{3} \) of its original resolution, and split each ray into four after every three marching steps, which is equivalent to doubling the resolution. After six steps, each pixel in the full resolution has a corresponding ray, which keeps marching until convergence.

**Aggressive Marching.** After the ray marching begins, we apply an aggressive strategy (Fig. 3 (b)) to speed up the marching process by updating the ray with \( \alpha \) times of the queried signed distance value, where \( \alpha = 1.5 \) in our implementation. This aggressive sampling has several benefits. First, it makes the ray march faster towards the surface, especially when it is far from surface. Second, it accelerates the convergence for the ill-posed condition, where the angle between the surface normal and the ray direction is small. Third, the ray can pass through the surface such that space in the back (i.e., SDF < 0) could be sampled. This is crucially important to apply supervision on both sides of the surface during optimization.

**Dynamic Synchronized Inference.** A naive parallelization for speeding up sphere tracing is to batch rays together and synchronously update the front end positions. However, depending on the 3D shape, some rays may converge earlier than others, thus leading to wasted computation. We maintain a dynamic unfinished mask indicating which rays require further marching to prevent unnecessary computation.

**Convergence Criteria.** Even with aggressive marching, the ray movement can be extremely slow when close to the surface since \( f \) is close to zero. We define a convergence criteria to stop the marching when the accuracy is sufficiently good and the gain is marginal (Fig. 3(c)). To fully maintain details supported by the 2D rendering resolution, it is sufficiently safe to stop when the sampled signed distance value does not confuse one pixel with its neighbors. For an object with a smallest distance of 100mm captured by a camera with 60mm focal length, 32mm sensor width, and a resolution of \( 512 \times 512 \), the approximate minimal distance between the corresponding 3D points of two neighboring pixels is \( 10^{-4} m \) (0.1mm). In practice, we set the convergence threshold \( \epsilon \) as \( 5 \times 10^{-5} \) for most of our experiments.

### 3.3. Rendering 2D Observations

After all rays converge, we can compute the distance along each ray as the following:

\[
d = \alpha \sum_{n=0}^{N-1} f(p^{(n)}) + (1 - \alpha) f(p^{(N-1)}) = d' + \epsilon, \tag{1}
\]

where \( \epsilon = (1 - \alpha) f(p^{(N-1)}) \) is the residual term on the last query. In the following part we will show how this computed ray distance is converted into 2D observations.

**Depth and Surface Normal.** Suppose that we find the 3D surface point \( \mathbf{p} = c + d\mathbf{v} \) for a pixel \((x, y)\) in the image, we can directly get the depth for each pixel as the following:

\[
z_c = \frac{d}{\sqrt{x^2 + y^2 + 1}}, \tag{2}
\]

where \((\tilde{x}, \tilde{y}, 1)^\top = K^{-1}(x, y, 1)^\top\) is the normalized homogeneous coordinate.

The surface normal of the point \( \mathbf{p}(x, y, z) \) can be computed as the normalized gradient of the function \( f \). Since \( f \) is an implicit signed distance function, we take the approximation of the gradient by sampling neighboring locations:

\[
\mathbf{n} = \frac{1}{2\delta} \begin{bmatrix} f(x + \delta, y, z) - f(x - \delta, y, z) \\ f(x, y + \delta, z) - f(x, y - \delta, z) \\ f(x, y, z + \delta) - f(x, y, z - \delta) \end{bmatrix}, \hat{\mathbf{n}} = \frac{\mathbf{n}}{||\mathbf{n}||}. \tag{3}
\]

**Silhouette.** The silhouette is a commonly used supervision for 3D shape prediction. To make the rendering of silhouettes differentiable, we get the minimum absolute signed distance value for each pixel along its ray and subtract it by the convergence threshold \( \epsilon \). This produces a tight approximation of the silhouette, where pixels with positive values belong to the background, and vice versa. Note that directly checking if ray marching stops at infinity can also generate the silhouette but it is not differentiable.

**Color and Semantics.** Recently, it has been shown that texture can also be represented as an implicit function parameterized with a neural network \([33]\). Not only color, other spatially varying properties, like semantics, material, etc,
Table 1. Ablation studies on the cost-efficient feedforward design of our method. The average time for each optimization step was tested on a single NVIDIA GTX-1080Ti over the architecture of DeepSDF [34]. Note that the number of initialized rays is quadratic to the image size, and the numbers are reported for the resolution of \(512 \times 512\).

<table>
<thead>
<tr>
<th>Method</th>
<th>size</th>
<th>#step</th>
<th>#query</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive sphere tracing</td>
<td>(512^2)</td>
<td>50</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>+ practical grad.</td>
<td>(512^2)</td>
<td>50</td>
<td>6.06M</td>
<td>1.6h</td>
</tr>
<tr>
<td>+ parallel</td>
<td>(512^2)</td>
<td>50</td>
<td>6.06M</td>
<td>3.39s</td>
</tr>
<tr>
<td>+ dynamic</td>
<td>(512^2)</td>
<td>50</td>
<td>1.99M</td>
<td>1.23s</td>
</tr>
<tr>
<td>+ aggressive</td>
<td>(512^2)</td>
<td>50</td>
<td>1.43M</td>
<td>1.08s</td>
</tr>
<tr>
<td>+ coarse-to-fine</td>
<td>(512^2)</td>
<td>50</td>
<td>887K</td>
<td>0.99s</td>
</tr>
<tr>
<td>+ coarse-to-fine</td>
<td>(512^2)</td>
<td>100</td>
<td>898K</td>
<td>1.24s</td>
</tr>
</tbody>
</table>

4. Experiments and Results

In this section, we first verify the efficacy of our differentiable sphere tracing algorithm, and then show that 3D shape understanding can be achieved through geometry based reasoning by our method.

3.4. Approximated Gradient Back-Propagation

DeepSDF [34] uses the conditional implicit function to represent a 3D shape as \(f_\theta(p, z)\), where \(\theta\) is the network parameters, and \(z\) is the latent code representing a certain shape. As a result, each queried point \(p\) in the sphere tracing process is determined by \(\theta\) and the shape code \(z\), which requires to unroll the network for multiple times and costs huge memory for back-propagation with respect to \(z\):

\[
\frac{\partial f_\theta}{\partial z} \bigg|_{z_0} = \alpha \sum_{i=0}^{N-1} \frac{\partial f_\theta(p^{(i)}(z), z)}{\partial z} \bigg|_{z_0} + \frac{\partial f_\theta(p^{(i)}(z_0), z_0)}{\partial p^{(i)}(z)} \bigg|_{z_0} \frac{\partial p^{(i)}(z_0)}{\partial z}.
\]

Practically, we ignore the gradients from the residual term \(e\) in Equation (1). In order to make back-propagation feasible, we define a loss for \(K\) samples with the minimum absolute SDF value on the ray to encourage more signals near the surface. For each sample, we calculate the gradient with only the first term in Equation (4) as the high-order gradients empirically have less impact on the optimization process. In this way, our differentiable renderer is particularly useful to bridge the gap between this strong shape prior and some partial observations. Given a certain observation, we can search for the code that minimizes the difference between the rendering from our network and the observation. This allows a number of applications which will be introduced in the next section.

4. Experiments and Results

In this section, we first verify the efficacy of our differentiable sphere tracing algorithm, and then show that 3D shape understanding can be achieved through geometry based reasoning by our method.
shows that a random initial code other spatially variant information. As an example, not only the signed distance function other than depth. With a pre-trained network encoding textured meshes, we can render high resolution color images under various resolution, camera viewpoints, and illumination.

Rendering Other Properties. Not only the signed distance function for 3D shape, implicit functions can also encode other spatially variant information. As an example, we train a network to predict both signed distance and color for each 3D location, and this grants us the capability of rendering color images. In Fig. 8, we show that with a 512-dim latent code learned from textured meshes as the ground truth, color images can be rendered in arbitrary resolution, camera viewpoints, and illumination. Note that the latent code size is significantly smaller than the mesh (vertices+triangles+texture map), and thus can be potentially used for model compression. Other per-vertex properties, such as semantic segmentation and material, can also be rendered in the same differentiable way.

4.2. 3D Shape Prediction

Our differentiable implicit SDF renderer builds up the connection between 3D shape and 2D observations and enables geometry based reasoning. In this section, we show results of 3D shape prediction from a single depth image, or multi-view color images using DeepSDF as the shape generator. On a high-level, we take a pre-trained DeepSDF and fixed the decoder parameters. When given 2D observations, we define proper loss functions and propagate the gradient back to the latent code, as introduced in Section 3.4, to generate 3D shape. This method does not require any additional training and only need to run optimization at test time, which is intuitively less vulnerable to overfitting or domain gap issues in pure learning based approach. In this section, we specifically focus on evaluating the generalization capability while maintaining high shape quality.

4.2.1 3D Shape Prediction from Single Depth Image

With the development of commodity range sensors, the dense or sparse depth images can be easily acquired, and several methods have been proposed to solve the problem of 3D shape prediction from a single depth image. DeepSDF [34] has shown state-of-the-art performance for this task, however requires an offline pre-processing to lift the input 2D depth map into 3D space in order to sample the SDF values with the assistance of the surface normal. Our differentiable render makes 3D shape prediction from a depth image more convenient by directly rendering the depth image given a latent code and comparing it with the given depth. Moreover, with the silhouette calculated from the depth map or provided from the rendering, our renderer can also leverage it as an additional supervision. Formally, we obtain the complete 3D shape by solving the following optimization:

$$
\arg\min_z \mathcal{L}_d(\mathcal{R}_d(f(z)), I_d) + \mathcal{L}_s(\mathcal{R}_s(f(z)), I_s),
$$

where $f(z)$ is the pre-trained neural network encoding shape priors, $\mathcal{R}_d$ and $\mathcal{R}_s$ represent the rendering function for the depth and silhouette respectively, $\mathcal{L}_d$ is the $L_1$ loss of depth observation, and $\mathcal{L}_s$ is the loss defined based on the differentiably rendered silhouette. In our experiment, the initial latent shape $z_0$ is chosen as the mean shape.
Table 2. Quantitative comparison between our geometric optimization with DeepSDF [34] for shape completion over partial dense and sparse depth observation on ShapeNet dataset [2]. We report the median Chamfer Distance on the first 200 instances of the dataset of [6]. We give DeepSDF [34] the groundtruth normal otherwise they could not be applied on the sparse depth.

<table>
<thead>
<tr>
<th></th>
<th>dense</th>
<th>50%</th>
<th>10%</th>
<th>100pts</th>
<th>50pts</th>
<th>20pts</th>
</tr>
</thead>
<tbody>
<tr>
<td>sofa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DeepSDF</td>
<td>5.37</td>
<td>5.56</td>
<td>5.50</td>
<td>5.93</td>
<td>6.03</td>
<td>7.63</td>
</tr>
<tr>
<td>Ours</td>
<td>4.12</td>
<td>5.75</td>
<td>5.49</td>
<td>5.72</td>
<td>5.57</td>
<td>6.95</td>
</tr>
<tr>
<td>Ours (mask)</td>
<td>4.12</td>
<td>3.98</td>
<td>4.31</td>
<td>3.98</td>
<td>4.30</td>
<td>4.94</td>
</tr>
<tr>
<td>plane</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DeepSDF</td>
<td>3.71</td>
<td>3.73</td>
<td>4.29</td>
<td>4.44</td>
<td>4.40</td>
<td>5.39</td>
</tr>
<tr>
<td>Ours</td>
<td>2.18</td>
<td>4.08</td>
<td>4.81</td>
<td>4.44</td>
<td>4.51</td>
<td>5.30</td>
</tr>
<tr>
<td>Ours (mask)</td>
<td>2.18</td>
<td>2.08</td>
<td>2.62</td>
<td>2.26</td>
<td>2.55</td>
<td>3.60</td>
</tr>
<tr>
<td>table</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DeepSDF</td>
<td>12.93</td>
<td>12.78</td>
<td>11.67</td>
<td>12.87</td>
<td>13.76</td>
<td>15.77</td>
</tr>
<tr>
<td>Ours</td>
<td>5.37</td>
<td>12.05</td>
<td>11.42</td>
<td>11.70</td>
<td>13.76</td>
<td>15.83</td>
</tr>
<tr>
<td>Ours (mask)</td>
<td>5.37</td>
<td>5.15</td>
<td>5.16</td>
<td>5.26</td>
<td>6.33</td>
<td>7.62</td>
</tr>
</tbody>
</table>

We test our method and DeepSDF [34] on 200 models on plane, sofa and table category respectively from ShapeNet Core [2]. Specifically, for each model, we use the first camera in the dataset of Choy et al. [6] to generate dense depth images for testing. The comparison between DeepSDF and our method is listed in Table 2. We can see that our method with only depth supervision performs better than DeepSDF [34] when dense depth image is given. This is probably because that DeepSDF samples the 3D space with pre-defined rule (at fixed distances along the normal direction), which may not necessarily sample correct location especially near object boundary or thin structures. In contrast, our differentiable sphere tracing algorithm samples the space adaptively with the current estimation of shape.

Robustness against sparsity. The depth from laser scanners can be very sparse, so we also study the robustness of our method and DeepSDF against sparse depth. The results are shown in Table 2. Specifically, we randomly sample different percentages or fixed numbers of points from the original dense depth for testing. To make a competitive baseline, we provide DeepSDF ground truth normals to sample SDF, since it cannot be reliably estimated from sparse depth. From the table, we can see that even with very sparse depth observations, our method still recovers accurate shapes and gets consistently better performance than DeepSDF with additional normal information. When the silhouette is available, our method achieves significantly better performance and robustness against the sparsity, indicating that our rendering method can back-propagate gradients effectively from the silhouette loss.

4.2.2 3D Shape Prediction from Multiple Images

Our differentiable renderer can also enable geometry based reasoning for shape prediction from multi-view color images by leveraging cross-view photometric consistency.

Specifically, we first initialize the latent code with a random vector and render a depth image for each of the input views. We then warp each color image to other input views using the rendered depth image and the known camera pose. The difference between the warped and input images are then defined as the photometric loss, and the shape can be predicted by minimizing this loss. To sum up, the optimization problem is formulated as follows,

\[
\arg\min_z \sum_{i=0}^{N-1} \sum_{j \in N_i} \|I_i - I_{j \rightarrow i}(R_d^i(f(z)))\|,
\]

where \(R_d^i\) represents the rendered depth image at view \(i\), \(N_i\) are the neighboring images of \(I_i\), and \(I_{j \rightarrow i}\) is the warped image from view \(j\) to view \(i\) using the rendered depth. Note that no mask is required under the multi-view setup.

Fig. 9 shows an example of the optimization process of our method. As can be seen, the shape is gradually improved while the loss is being optimized.

We take PMO [24] as a competitive baseline, since they also perform deep learning based geometric reasoning via optimization over a pre-trained decoder, but use the triangular mesh representation. Their model first predicts an initial mesh from a selected input view and improve the quality via cross-view photo-consistency. Both the synthetic and real datasets provided in [24] are used for evaluation.

In Table 3, we show quantitative comparison to PMO on their synthetic test set. It can be seen that our method achieves comparable results with PMO [24] from only random initializations. Note that while PMO uses both the encoder and decoder trained on the PMO training set, our DeepSDF decoder was neither trained nor finetuned on it. Besides, if the shape code for PMO, instead of being predicted from their trained image encoder, is also initialized randomly, their performance decreases dramatically, which indicates that with our rendering method, our geometric rea-
Figure 10. Robustness of geometric reasoning via multi-view photometric optimization. (a) Performance w.r.t changes on camera focal length. (b) Performance w.r.t noise in the initialization code. Our model is robust against focal length change and not affected by noise in the latent code since we start from random initialization. In contrast, PMO is very sensitive to both factors, and the performance drops significantly when the testing images are different from the training set.

Generalization Capability To further evaluate the generalization capability, we compare to PMO on some unseen data and initialization. We first evaluate both methods on a testing set generated using different camera focal lengths, and the quantitative comparison is in Fig. 10 (a). It clearly shows that our method generalizes well to the new images, while PMO suffers from overfitting or domain gap. To further test the effectiveness of the geometric reasoning, we also directly add random noise to the initial latent code. The performance of PMO again drops significantly, while our method is not affected since the initialization is randomized (Fig. 10 (b)). Some qualitative results are shown in Fig. 11. Our method produces accurate shapes with detailed surfaces. In contrast, PMO suffers from two main issues: 1) the low resolution mesh is not capable of maintaining geometric details; 2) their geometric reasoning struggles with the initialization from image encoder.

We further show comparison on real data in Fig. 12. Following PMO, since the provided initial similarity transformation is not accurate in some cases, we also optimize over the similarity transformation in addition to the shape code. As can be seen, both methods perform worse on this challenging dataset. In comparison, our method produces shapes with higher quality and correct structures, while PMO only produce very rough shapes. Overall, our method shows better generalization capability and robustness against domain change.

5. Conclusion

We propose a differentiable sphere tracing algorithm to render 2D observations such as depth maps, normals, silhouettes, from implicit signed distance functions parameterized as a neural network. This enables geometric reasoning in 3D shape prediction from both single and multiple views in conjunction with the high capacity 3D neural representation. Extensive experiments show that our geometry based optimization algorithm produces 3D shapes that are more accurate than SOTA, generalizes well to new datasets, and is robust to imperfect or partial observations. Promising directions to explore using our renderer include self-supervised learning, recovering other properties jointly with geometry, and neural image rendering.

Acknowledgements

This work is partly supported by National Natural Science Foundation of China under Grant No. 61872012, National Key R&D Program of China (2019YFF0302902), and Beijing Academy of Artificial Intelligence (BAAI).
References


[29] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks:


