

Z2P: Instant Visualization of Point Clouds

G. Metzger¹, R. Hanocka⁴, R. Giryas¹, N. J. Mitra^{2,3}, D. Cohen-Or¹

¹Tel Aviv University
²University College London
³Adobe Research
⁴University of Chicago

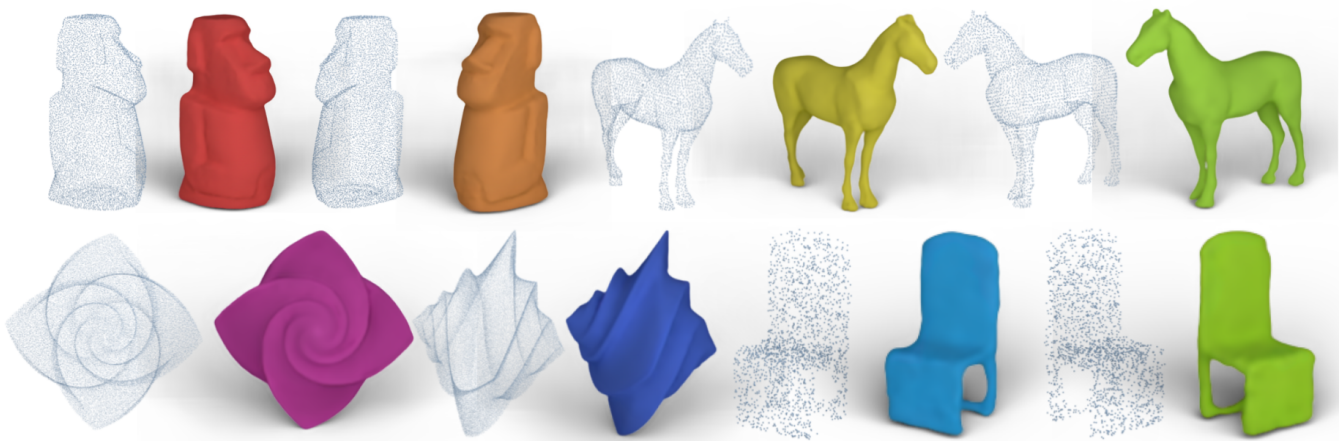


Figure 1: Point cloud inputs and resulting visualization using our Z2P, conditioned on different colors and lighting positions.

Abstract

We present a technique for visualizing point clouds using a neural network. Our technique allows for an instant preview of any point cloud, and bypasses the notoriously difficult surface reconstruction problem or the need to estimate oriented normals for splat-based rendering. We cast the preview problem as a conditional image-to-image translation task, and design a neural network that translates point depth-map directly into an image, where the point cloud is visualized as though a surface was reconstructed from it. Furthermore, the resulting appearance of the visualized point cloud can be, optionally, conditioned on simple control variables (e.g., color and light). We demonstrate that our technique instantly produces plausible images, and can, on-the-fly effectively handle noise, non-uniform sampling, and thin surfaces sheets.

1. Introduction

Point clouds are a popular and flexible representation of 3D shapes. A slew of works have successfully employed deep neural networks to synthesize point clouds, for shape synthesis [ADMG18; LZZ*18; YHH*19; HHGC20; GBZC20], upsampling [YLF*18b; YWH*19], consolidation [YLF*18a; MHGC20], shape completion [YKH*18], denoising [RLG*20], among others. Thus, generating an increasing demand for a fast and effective technique to visualize point clouds. Yet, neurally synthesized point clouds do not contain a globally consistent normal orientation (since

standard loss functions do not trivially enable normal regression [MHZ*21]), which is a prerequisite for rendering them using surface reconstruction.

Since points are zero-dimensional entities, they cannot be rendered directly. One approach enables point cloud visualization by converting each point into a local tangent patch (i.e., splats) [ZPVG01; BK03]. Since point samples are irregular and unstructured, different heuristics for deciding the size of local splats inevitably lead to holes and overlaps. Moreover, these techniques often assume normals are provided as input, or ought to be esti-

mated on-the-fly [DB07; PJW12]. One approach for avoiding the difficult normal estimation problem is through screen space operators [PGA11], at the expense of limited shading and visualization abilities.

A more meticulous approach is to first reconstruct a mesh surface from the input point samples [KBH06; KH13; HMGC20] and then render the mesh. However, reconstructing a surface is a notoriously difficult problem. Such an approach appears to be *overkill* if the end goal is to simply visualize the point set. Further, it inherits the problems of any reconstruction algorithm. For example, neural synthesized point clouds lack normal orientation and often contain interior points, resulting in poor reconstruction results. Moreover, surface reconstruction is unstable and not well defined in the case of thin surfaces and sheets. Finally, these methods are often slow, which is not feasible for a quick preview visualization. One exception is [KTB07], which allows for faster reconstruction without normal information, but has limited success on sparse point clouds (see Figure 8).

We cast the problem as an image-to-image translation problem whereby z-buffers, *i.e.*, depth-augmented point features as viewed from target camera view, are directly translated by a neural network to rendered images, conditioned on control variables (e.g., color, light). Once trained, our network produces visualization using a single forward pass, and hence is instantaneous. Further, we support scene controls, such as light position and colors as an input condition to the network using adaptive instance normalization. The convolutional nature of the network enables learning a mapping at the patch level, which is location invariant. Yet, the shadows in the scene do depend on the location. Thus, by leveraging a positional encoding, we break this invariance and allow the shading effects to depend on the location.

Our network is trained on automatically generated pairs of point clouds and rendered meshes, and only requires training on a few unique meshes to achieve generalization. We generate a copious amount of paired data from each mesh, by applying random lighting, color, and rotation augmentations. During inference our method can generate plausible images from point clouds where the ground-truth is unknown (*i.e.*, noisy point clouds from neural networks or scanning devices), bypassing the challenging reconstruction or point normal orientation problem altogether. Moreover, it is significantly faster than the rendering engine it was trained on: our preview visualization takes under one second, while the full-blown rendering engine takes over 10 seconds.

We show that our method can act as an effective, quick-and-dirty point cloud preview, while still providing useful control. We demonstrate that our approach can successfully handle noise, non-uniform sampling, and in particular thin surfaces and sheets (see Figure 9). Our experimental evaluations show that our technique is comparable or better than existing techniques, and is considerably faster.

2. Related Work

Splatting. The most common and direct way to render point clouds is through splatting, which are small planes placed at each point in the input point cloud, and point in the normal direction [ZPVG01;

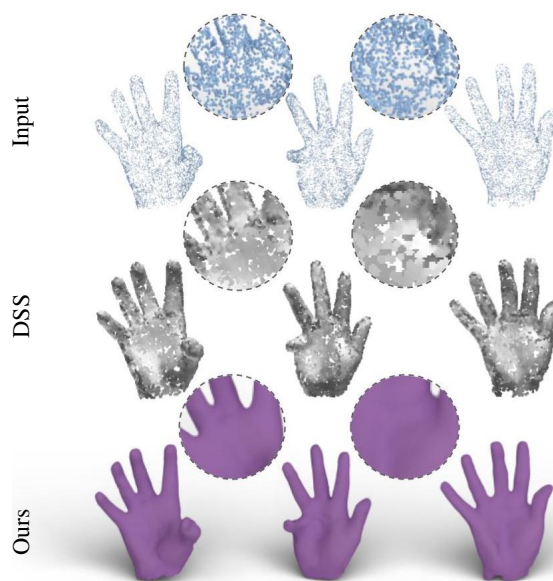


Figure 2: Our method can directly visualize point clouds with non-uniform sampling. Note the self shadow effects on the palm of the left example.

DB07; PJW12]. Splatting methods generally require normals as input in order to set the splat direction. Some splatting methods such as [DB07; PJW12] can estimate the normal direction in screen space, instead of relying on explicit normals as input. Both [DB07; PJW12] are able to work in real-time and utilize a GPU. The orientation of the normal vector can also be considered for shading purposes, to detect back facing points. Results of different splatting methods can be seen in Figures 2, 8, 15. Splatting-based methods are adequate for fast and simple visualizations, but may be limited in some features as described in Table 1.

Screen Space Operators. PINTUS et al. [PGA11] developed a screen space technique that is able to render point clouds in real-time without normals. To render the point clouds, the points are first projected onto a texture, then the method detects the visible points and interpolates the depth values over the rest of the screen. Screen space deferred shading is applied to highlight surface curvature. Although this method is real-time and does not require normals, it can only perform deferred shading.

Surface reconstruction. Reconstructing a surface mesh from point clouds is a long standing problem in computer graphics that has been studied extensively for over 20 years [BTS*17]. A common approach triangulates a set of points [EM94; BMR*99]. Alternatively, an implicit function can be built from the point cloud and normals, and the surface mesh is extracted by finding the zero-crossings [HDD*92; KBH06; KH13]. Surface reconstruction can also be obtained using a mesh deformation technique, where an initial mesh is incrementally displaced to fit an input point cloud [SLS*06; HMGC20].

A common prerequisite for surface reconstruction is calculating a globally consistent normal orientation [MN03; LW10; GKOM18;

SSG17; JBG19]. Though some techniques, such as [MDD*10; HMGC20] can reconstruct surfaces with unoriented normals. Triangulation techniques (such as ball-pivot [BMR*99]), handle thin sheets well, yet inevitably lead to undesirable holes in the reconstruction. On the other hand, implicit surface reconstruction (such as Poisson [KBH06; KH13]) tend to handle holes better, yet occupancy is not-well defined in the case of sheets. Finally, this route is computationally expensive since it requires both reconstructing a surface and then rendering an image.

KATZ et al. [KTB07] developed a method to detect visible points from a raw point cloud without normals. As a by product of the hidden point removal (HPR), the method also allows for a "quick-and-dirty" single view reconstruction *i.e.* a partial mesh that constrains a surface only from the view direction. A comparison to KATZ et al. [KTB07] can be found in Figure 8 and in the supplementary material.

Deep rendering. Since the introduction of the rendering equation [Kaj86], many works have proposed techniques to speedup the rendering process, for example, using *deferred shading* [ST90] to produce a render from 2D G-buffers (*e.g.*, depth image, normal maps, etc).

Although deep rendering works and our work both aim at producing an image from a 3D object from a single view, rendering works have a different focus. Rendering methods usually assume a high quality description of the 3D geometry that can be queried, like continuous surfaces, and focus on producing highly realistic visualizations while supporting controls like lighting, materials, etc. Our work, on the other hand, enables a quick preview of the shape (bypassing the surface reconstruction problem), which does not qualify as a full fledged rendering engine. Recently, deep learning has been used to perform deferred shading [NAM*17] which calculates the G-buffer from the complete 3D scene. However, estimating these maps (*e.g.*, occlusions, light and normal information) is non-trivial for point clouds. RenderNet [NLBY18] learns to produce 2D images from a 3D voxel representation of the shape. Adopting this approach to point clouds is wasteful as it requires creating a voxelized volume (reducing the resolution down to $64 \times 64 \times 64$), as well as require oriented normals. HERMOSILLA et al. [HMRR19] learns to shade surfaces using deep neural networks and G-buffers, which is a faster alternative than path-tracing. Note that we do not have access to G-buffer information in our case.

Neural point cloud rendering has been studied in [KSL20; DZL*20; AUL19]. These works mainly focus on rendering novel views from a single scene using a sparse set of input views. In this setting, since a descriptor is optimized for each point, it does not extend to point clouds outside of the training set (single scene). This type of problem is different than ours, as we do not aim at visualizing a single scene, instead our technique is generalize to unseen point clouds. Moreover, our system achieves complete disentanglement and control of lighting, color, and ambient shadows, as compared to these works that receive shading and color information from the given RGB images of the scene. A comparison between our work and alternative methods is depicted in Table 1. Recently, [CCCM21] proposed a controllable neural rendering pipeline for neural 3D shapes represented by neural implicit.

An orthogonal line of works propose differential render-

ing [YSW*19; CGL*19] techniques for a plug-and-play module in deep learning architectures. Different than the objective of this work, these works are used to perform shape reconstruction or inverse rendering (*i.e.*, recovery of unknown 3D scene and lighting).

3. Method

We formulate the visualization task as an image-to-image translation from z-buffers of projected point clouds to *surface* images, conditioned on the visualization settings. Our framework is simple, yet effective. We automatically generate simulated training data to train a fully-convolutional network to learn a mapping from point cloud z-buffers to rendered mesh images. We inject visualization control through Adaptive Instance Normalization (AdaIN) [HB17] layers, enabling a disentangled representation of visualization parameters and shape.

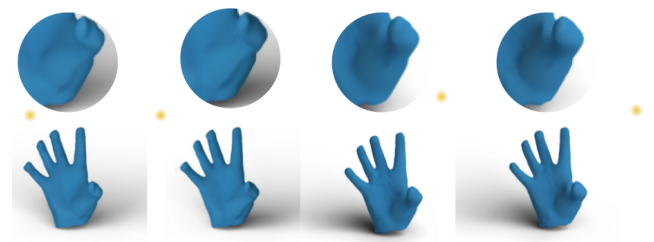


Figure 3: Self shadow effects are plausibly generated with our method. Light position is visualized in yellow.

Synthesizing shadows requires long range interactions between pixels (see Figures 3 and 7). This is achieved by the U-Net [RFB15] structure which takes into account various scales of the input. Yet, in order to enable the network to effectively process the light direction, we break the translational symmetry of the convolutions by appending positional encoding to the input. During inference, we use real point clouds (*i.e.*, noisy point clouds obtained from neural networks or scanning devices) to obtain a plausible quality image, previewed in under a second, bypassing the challenging surface reconstruction or point normal orientation problem altogether.

3.1. Point Cloud Projection

We use a 2D z-buffer projection of the point cloud as input to our network. XIE et al. [XWZ*21] demonstrated a similar projection was useful for their adversarial loss. Each point in the point cloud p is projected onto the nearest 2D image coordinate (u_p, v_p) , with an intensity value proportional to the depth d_p , *i.e.*, distance to the image plane. The projection is a perspective transformation according to the camera pinhole model with extrinsic (camera location) and intrinsic (*e.g.*, focal length) parameters. Pixels in the 2D image plane that do not correspond to any 3D points receive an intensity of 0, while other pixels receive an intensity value according to

$$z(u_p, v_p) = e^{-(d_p - \alpha)/\beta}, \quad (1)$$

where α and β are hyperparameters that we fix once throughout training, inference and all experiments. The width of the intensity

	Ray Tracing	Splatting	Screen Space	Reconstruction	NPR	Ours
Raw xyz points	✗	✓	✓	✓	✗	✓
Handles points w/o normals	-	✓/✗	✓	✗	✓	✓
Generalizes to unseen objects	✓	✓	✓	✓	✗	✓
Robust to internal points	-	✓	✓	✗	✓	✓
Robust to number of points	-	✗	✓	✗	✗	✓
Shadows	✓	✗	✗	✓	-	✓
Per point colors	-	✓	✓	✓	✓	✗
Scene level scans	-	✓	✓	✓	✓	✗

Table 1: Comparing Z2P to other visualization techniques. Reconstruction refers to reconstructing a mesh and rendering it with ray tracing, and neural point rendering (NPR) refers to the prior works of [KSL20; DZL*20; AUL19]. Splatting contains ✓/✗ for requiring normals, as some works such as [DB07; PJW12] are able to estimate normals on the fly, and therefore technically do not require normals as input. Screen Space refers to the work of [PGA11], as discussed in the related work it should be noted that this work is limited to deferred shading.

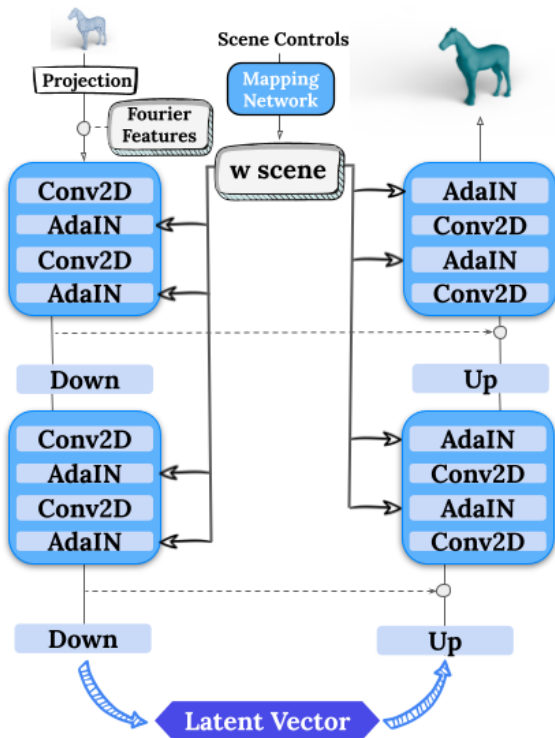


Figure 4: Overview. We project the point cloud into a 2D z buffer image, which is the input to our network. We append positional encoding (Fourier features) in order to model long-range effects required to generate shadows. We inject scene controls into AdaIN after each convolutional layer, enabling control of the light and color in the scene.

values for each points' 2D projection on the image is within a 5×5 window. In the case where multiple points are assigned to the same pixel, we use only the closest one to determine the intensity.

The intuition behind our z-buffer image is that points far away receive a value approaching the background intensity, whereas close points are exponentially brighter than occluded ones. This makes

2D features consistent with the visualization requirements, since far points are usually occluded, which should result in an intensity value that is close to background and much lower than points which are in front of it. Note that our z-buffer image may contain back-facing points, since we do not use oriented point normals. Despite this, we observe that our network is able to cope with such hidden surfaces well.

3.2. Training Data Generation

Our training data is automatically generated using Blender 3D engine [Com21], which provides explicit control over the rendering parameters. We sample meshes to produce corresponding pairs of point cloud and mesh data. In order to generate a large amount of geometric variety using only a small number of unique meshes, we apply random rotations and create different point cloud instances from the same mesh. Note that due to the convolutional nature of the network, which shares weights across patches of the input z-buffer, the *effective number* of training examples is larger than the number of pairs of input images.

For each pair of point cloud and underlying mesh pair, we randomly select a color and light location which is used to automatically render an image using Blender Cycles engine. The light is modeled using a flat emission surface with a floor plane for catching shadows and ambient light.

This results in many pairs of training inputs (z_i, s_i) and outputs r_i , where z_i is the 2D projection of the point cloud, s_i is a vector containing the settings (color and light position), and r_i is the corresponding blender rendered image.

The training data is composed of 20 unique meshes each augmented 400 times with rotation, color, and lighting position. Each mesh has 10 corresponding point clouds that were sampled from it. This results in 80K pairs of training examples.

3.3. Network Architecture

We formulate our visualization task as an image-to-image translation from z-buffers to rendered images (with an alpha channel for matting). However, this is not a simple image-to-image setup as the target visualization is conditioned on the given color and

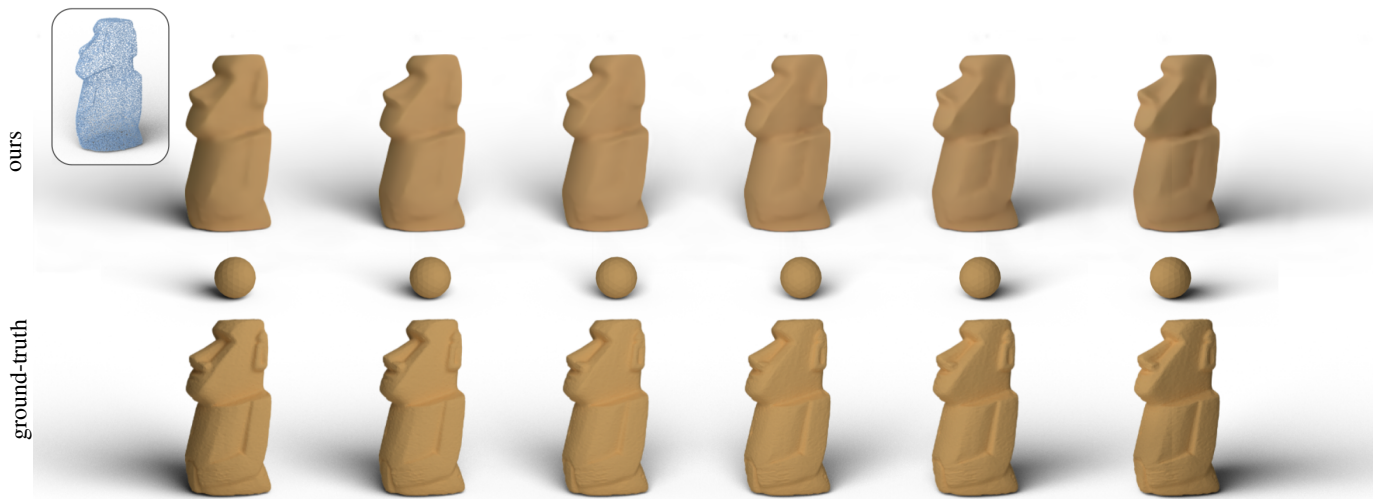


Figure 5: Lighting control. Top: our visualized result conditioned on slowly changing light position produces a smooth transformation of the shade and objects lighting. Bottom: Ground truth mesh renderings under corresponding conditions. The shaded sphere helps visualizing the light direction.



Figure 6: Color control. We keep a constant light and view angle, while continuously varying the input color.

light position. To enable better control, we propose a modified U-Net [RFB15] architecture where we add AdaIN layers to control the color and light position, and use positional encoding to produce accurate shadows. Below, we describe each of these added components. In Section 4, we show the importance of these components through an ablation study.

Visualization control. Our architecture enables independent control of both the object color as well as light position of the visualization. This disentanglement can be seen in figures 6 and 5. We inject the color to the network as a 3-dimensional vector, which defines the RGB values of the target shape. The light position is given as a 3-dimensional vector corresponding to the light position in spherical coordinates with respect to the camera. The two vectors are concatenated to produce the settings vector $s \in \mathbb{R}^6$. We map the settings vector $s \in \mathbb{R}^6$ to a higher dimensional vector $w \in \mathbb{R}^{512}$ by an MLP mapping network. The vector w is used to produce the mean and standard deviation that are used by the AdaIN layers. Notice that while w is shared across all AdaIN layers, each of them uses a different (learned) affine transform to control its parameters.

This dependence on the light and color through AdaIN allows controlling the scene setting.

Adaptive Instance Normalization (AdaIN). The original U-Net uses batch normalization after each convolution operation. However, to control the color and shadows of the resulting image, we replace batch normalization with AdaIN layers. These layers normalize the feature map from the prior convolution operation. This was also shown to be an effective technique for controlling global style effects in [KLA19].

Each AdaIN layer ADA_i has a target feature length of size f_i . In order to use the global vector w , each AdaIN layer contains two additional affine transformations $A_{\beta,i}$, $A_{\gamma,i}$ that receive an input of size 512 and output a vector of size f_i .

Formally, each AdaIN layer takes as input a feature map x of size $H \times W \times f_i$ and the vector w . The affine transformation is used to calculate $\beta_i = A_{\beta,i} \cdot w$ and $\gamma_i = A_{\gamma,i} \cdot w$, and the feature map x is normalized according to Equation 2 below, where $\mu(x)$ and $\sigma(x)$ are the mean and standard deviation over the spatial dimensions per each channel,

$$\hat{x} = \gamma_i \left(\frac{x - \mu(x)}{\sigma(x)} \right) + \beta_i. \quad (2)$$

Positional encoding. U-Net is a fully convolutional network which means it has a limited receptive field, and is translation equivariant. The light position is used to control shadows. This requires long range interactions between pixels, and should not be translation equivariant. For example, when lifting an object in the vertical direction, shadows should remain at the bottom of the image, as the object moves farther away from the bottom floor plane (see Figure 7).

To allow for long interactions and break the translational symmetry, we append positional encoding features to the input image

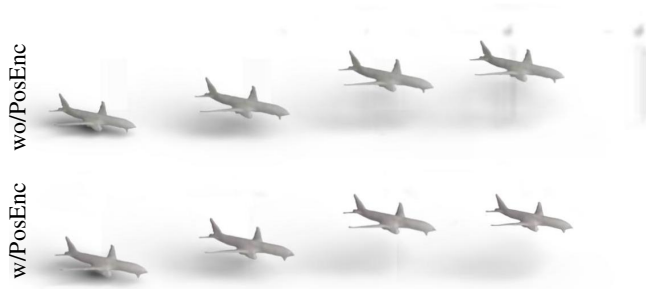


Figure 7: Shadow effects require positional encoding. During inference, we move the object location upwards (without having trained on this type of data). Observe how the positional encoding correctly learns the shadow interactions, keeping the floor plane in the correct location.

that depend on the pixel location. This enables the convolution filters to grasp the global context. We use random Fourier features, which has been successful at modeling both short and long interactions [TSM*20].

Formally, each pixel gets a normalized coordinate value (u_i, v_i) in the range $[0, 1]$. At the beginning of training, ten $\{\omega_j\}$ values are sampled from a uniform distribution $\omega_j \sim U[0, 10]$, to produce 40 features for each pixel of the form:

$$enc_i = [\{\sin(\omega_j \cdot u_i)\}_j, \{\sin(\omega_j \cdot v_i)\}_j, \{\cos(\omega_j \cdot u_i)\}_j, \{\cos(\omega_j \cdot v_i)\}_j] \quad (3)$$

Each encoding enc_i is concatenated to its corresponding pixel in the z-buffer, which creates additional input features.

3.4. Losses

Our network is trained using a combination of simple reconstruction losses. The network produces a 4-dimensional image, three color channels (for RGB) as well as an alpha channel for transparency. We use an L_2 loss between the network-predicted color image and the ground-truth color image. We also calculate the magnitude of the three color channels (pixel-wise) for both images and calculate the L_1 -norm between them. We also compute the L_1 distance between the ground truth and predicted alpha map intensities.

4. Experiments

We performed various experiments to validate our network performance as well as compare to existing surface reconstruction techniques. We demonstrate the ability to generalize on point clouds, where the ground-truth surface is unknown, such as real scans, as well as on point clouds that were synthesized from neural networks. We justify various components of our network architecture and perform run-time comparisons. All results in this paper are shown on held out validation or test set examples. There are extended results and evaluations in the supplementary material and video.

4.1. Visualization Control

We control the visualized result by modifying different settings: light position, color, and viewing angle. For example, in Figure 6 (more in supplemental) we control the color of the visualized result. In Figure 5 we move the light position in the horizontal direction, which changes the ambient shading on the object as well as the shadows on the ground (yet the color is held constant). We also show the visualization result from multiple view points in Figure 10, while holding light position and color constant. See supplemental video for temporal animations of modifying different visualization controls, as well as the supplemental material.

4.2. Handling Noisy Inputs

Depending on the use case, one may want to disregard noise, even if it is part of the input. We trained our model with examples containing uniform noise to produce a type of *denoising* visualizations. To test the ability to disregard noise, we added an increasing level of Gaussian noise during inference (different than the noise we trained on), in Figure 8. The last row contains the results on Z2P trained with uniform noise, and the second to last row is Z2P trained on

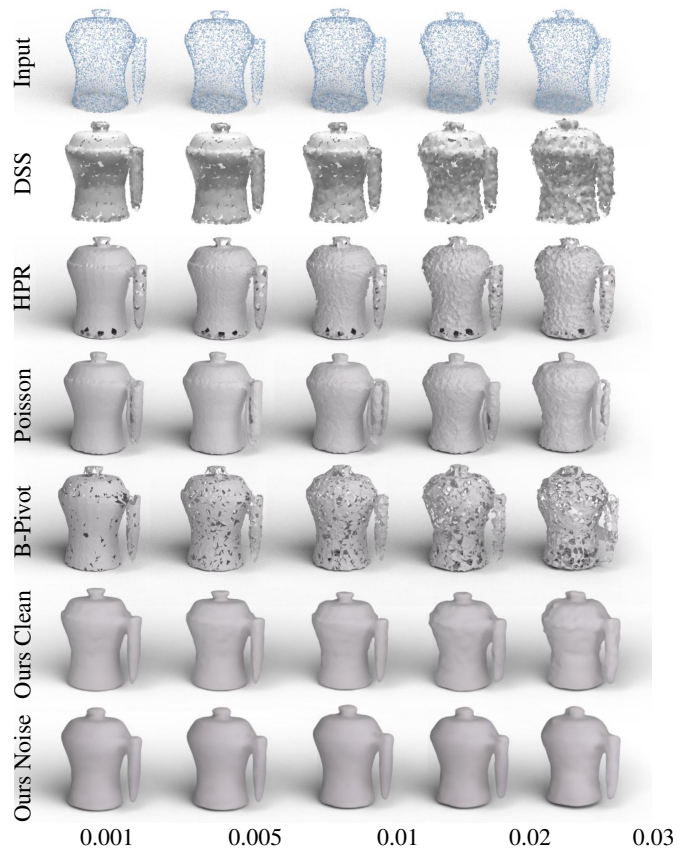


Figure 8: Handling noisy inputs. We compare against different approaches for visualizing a noisy point cloud with increasing levels of Gaussian noise (standard deviation at the bottom). The last row shows results of our network trained with examples corrupted with uniform noise.

clean data. The noise-removal version visualizes results that are noticeably smoother than the baseline Z2P, however this may come at the cost of over-smoothing fine details, since they can be attenuated in the presence of noise.

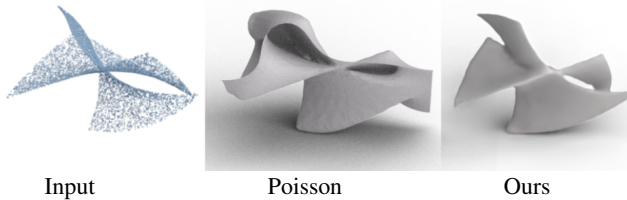


Figure 9: Sheets are especially hard to orient, reconstruct, and may be unprintable. Popular techniques such as Poisson reconstruction [KBH06] are not well defined in these cases (due to the open surface and poor normals). Our method, which operates on raw point clouds, can handle such cases as it does not rely on oriented normals or other surface attributes.

4.3. Robustness To Sampling

We show robustness to different types of sampling methods in Figure 11. Note, that despite only being trained on uniformly sampled meshes, our method is able to generalize to different sampling methods, as well as noise from neural synthesized point networks and real scanning devices.

We also evaluate the ability of our method to cope with point clouds at different scales, which results in a different point density on the z-buffer image in the supplemental material.

4.4. Generalization

We demonstrate the ability of our technique to visualize images from point clouds where the ground-truth underlying surface is unknown. Such an example is point clouds synthesized from neural networks. These are interesting since they are challenging for surface reconstruction methods [MHZ*21], since they do not contain normal information and often have inner points and single sheets. In Figure 15 and in the supplementary, we show our ability to robustly handle these challenging cases, without oriented normals and with the unknown noise distribution. Future deep learning approaches may use our work to better visualize the generated shapes.

We also compared against Point2Mesh [HMGC20] which also does not require normal orientation, in Figure 12. Additional comparisons can be found in the supplementary.

Scans acquired from sensor devices, are usually obtained as a single occluded view. These scans are not well defined for Poisson reconstruction, but can be used with an explicit triangulation method like ball pivot (which tends to introduce noise). Figure 13 shows results of a single view scan obtained from a low-end Intel RealSense SR300 scanner. We also show results on a lidar scanner, which contains an open surface which is especially challenging for surface reconstruction techniques, such as Poisson 14.

4.5. Timing Comparison

Table 2 shows the run-time of each individual component in our method. The table shows the timings for: projecting the point cloud onto the 2D coordinate plane (*project*), calculating the z-buffer intensity (*z-buffer*), forward pass through the network (*forward*), and the average overall time it takes for one example (*avg per example*). Visualizing a point cloud from start to finish takes under three seconds for a point cloud of size 5k, where the most expensive computation is the z-buffering process. Since the z-buffering code is written in Python, this part can be significantly accelerated using GPU and C++. Training our method takes around 24 hours on a single GTX 1080 Ti GPU. Rendering the train and validation set takes 48 hours (using a single GPU with Blender).

batch size	project ms	z-buffer sec	forward ms	avg per example sec
1	<0	2.60	20	2.62
2	10	5.04	40	2.54
3	10	7.88	60	2.65
4	10	10.70	80	2.69
5	20	13.28	110	2.68
6	20	17.48	120	2.93

Table 2: Run time break down. Visualizing a point cloud from start to finish using our method takes under 3 seconds. The z-buffer operation takes almost all the time, which is written in unoptimized Python code.

Rendering an image from point clouds by first estimating a reconstruction is computationally expensive. As shown in Table 3, Reconstruction requires estimating the point normals and propagating their orientation (~ 3 seconds), running the reconstruction algorithm itself (~ 1 or ~ 10 seconds), and then rendering the result with blender (~ 10 seconds). In total, Poisson reconstruction and ball pivot take 13.85 and 23.92 seconds, respectively, whereas our technique on the same point cloud takes less than 3 seconds.

steps [sec]	Poisson	B-pivot
normal estimation	2.76	2.76
reconstruction	10.63	1.03
render	10.53	10.06
total time	23.92	13.85

Table 3: Run time break down for surface reconstruction approaches. In total, Poisson reconstruction and ball pivot take 13.85 and 23.92 seconds, respectively, whereas our technique on the same point cloud takes less than three seconds (Table 2).

4.6. Architecture Ablation

The design choices of our architecture are explained through several experiments. We control the visualization through AdaIN, which has been shown to be a powerful tool for controlling style parameters. We compare it to a feature-baseline, where instead of injecting the visualization settings through AdaIN, we create a variant of our network which appends the visualization settings as extra



Figure 10: Point clouds can be rendered from multiple views while producing a coherent result with respect to both color and shadows.

input features (repeated for every input z-buffer pixel). Note how in Figure 16, the AdaIN version maintains a consistent color regardless of sampling density. By injecting controls at a global scale with AdaIN, which only changes second order feature map statistics, we obtain a framework with better style consistency and control. We also test the impact of the positional encoding. Figure 7 shows that it is necessary for correctly modeling the long range interactions needed to generate shadows.

4.7. Splatting

In Figures 2, 15, 16 and the Supplementary Material we compare to splatting techniques. Specifically, we compare to two splat-

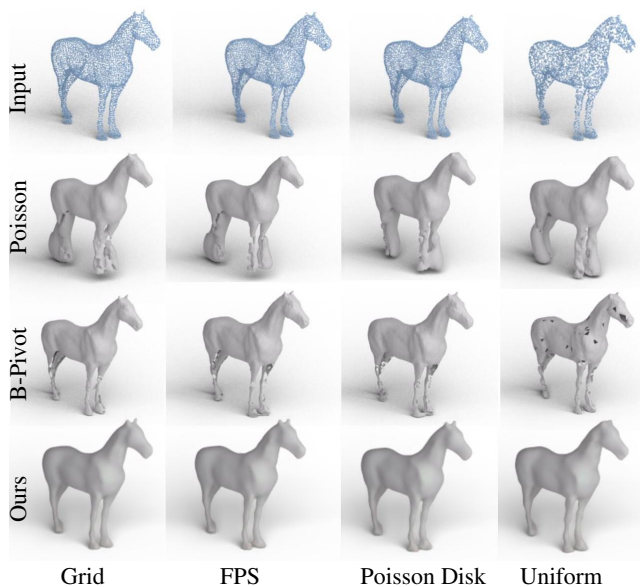


Figure 11: Robustness to different point sampling methods. Poisson reconstruction and ball-pivot (top, middle rows respectively) vary in quality across different sampling methods, our method is able to produce a high quality consistent visualization of the point cloud.

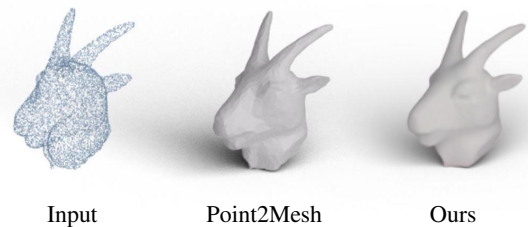


Figure 12: Comparison to [HMGC20]. Though Point2Mesh also produces good results, it requires a long time to converge (30-60 minutes), compared to our result that is obtained in seconds.

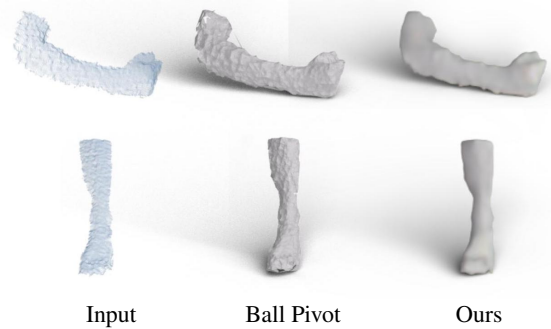


Figure 13: Results of visualizing single view point clouds obtained from Real Sense depth scanner.

ting methods, calculated using MeshLab [CCC*08] and using DSS [YSW*19]. We use the forward renderer from DSS, which can be seen as splatting using ZWICKER et al. [ZPVG01]. All splats are rendered with double-sided shading, which is oblivious to the normal orientation.

As the point clouds used in the paper are relatively sparse, holes are visible in the splatting visualizations. The sparsity also affects normal estimation, as it is calculated using local neighborhood information, and causes noisy *speckle* artifacts, as seen in Figure 15.

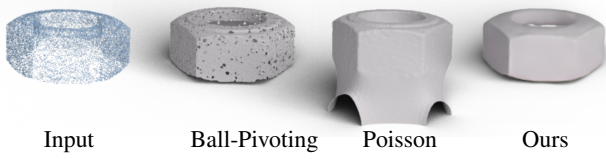


Figure 14: Results on real lidar scanned data, which contains holes, noise and open surfaces.

Splating techniques are better designed for dense point sets, as can be seen in Figure 16, as well as when the points already contain shaded color information.

4.8. Material Control

All the figures presented throughout the paper, are the results of a model trained with the *default* Blender diffuse material, which is sufficient for plausible visualization of point clouds. To allow for more control over the appearance, we trained our method on a spe-

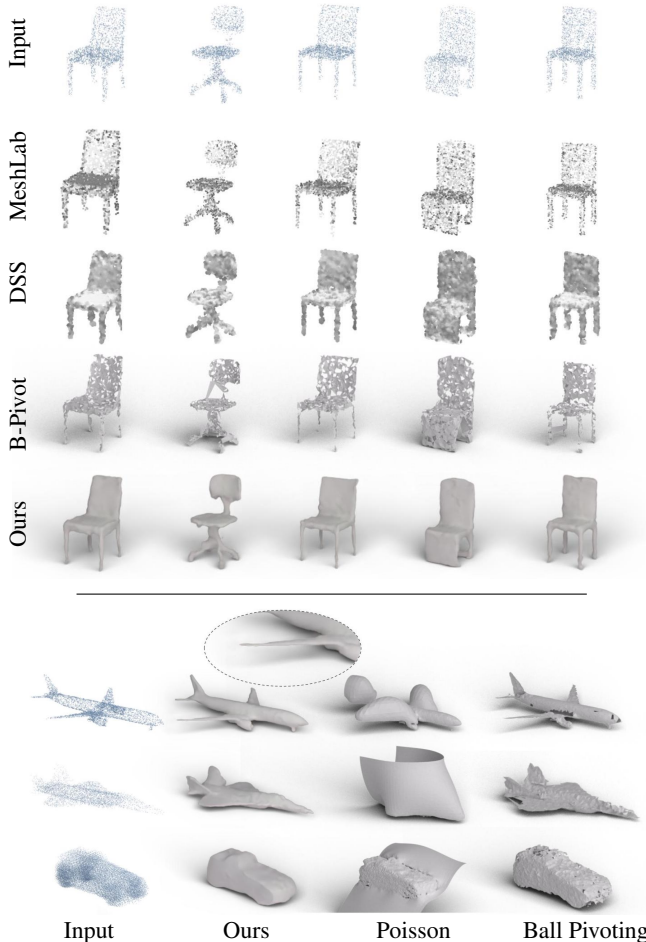


Figure 15: Our results on different chairs, planes, and cars generated by [CYA*20]. Poisson reconstruction was not able to produce a meaningful result for the chairs with thin surfaces in the top row.

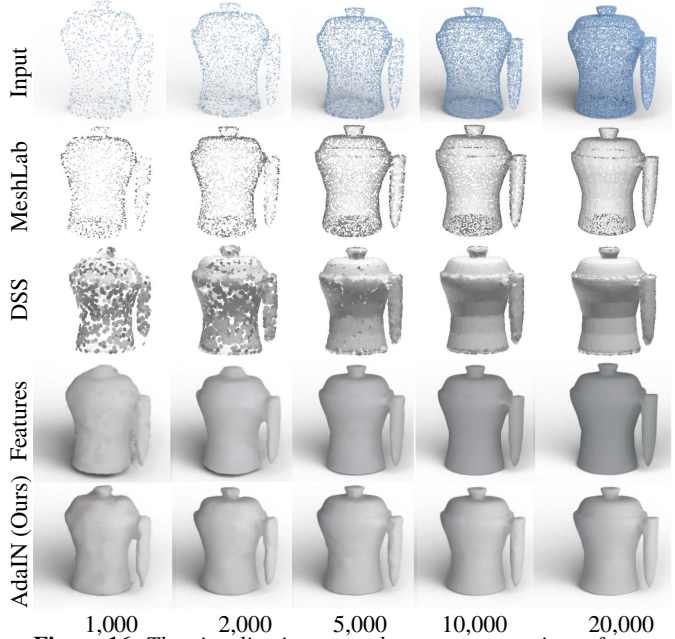


Figure 16: The visualization control vector as extra input features leads to an inconsistent lighting color result with respect to the input density.

cial dataset that uses the Blender *Principled BSDF* shader. In our experiment, we trained the network to control over the dominant *Metallic* and *Roughness* attributes.

In this special dataset, for each training example we sampled random color and light direction controls, as in the original dataset, together with random *Metallic* and *Roughness* values. Similar to the basic training method, we input the original controls together with the two additional (*Metallic* and *Roughness*) controls to the network through the AdaIN layers.

Figure 17 shows two examples with different *Metallic* and *Roughness* values as input. The supplementary material contains

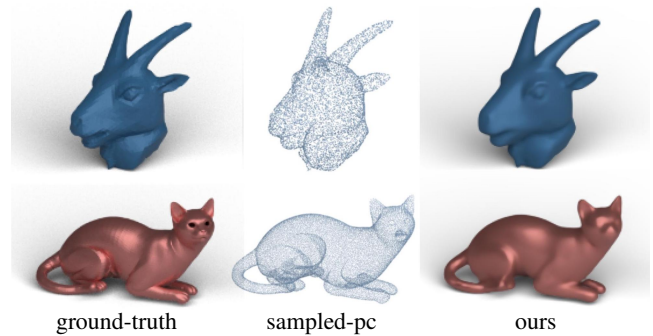


Figure 17: Two examples of the results of our method, trained on a dataset with the *Principled BSDF* shader. The examples shows that our method can control *Roughness* and *Metallic* attributes. More examples can be seen in the supplementary material.

a sweep over many *Metallic* and *Roughness* values and the corresponding outputs for a better evaluation.

5. Conclusions and Future Work

We have presented a neural visualization technique for point clouds. We refer to this technique as *instant visualization* as we bypass the cumbersome step of explicitly converting the point cloud to a continuous surface prior to rendering. Moreover, our instant visualization relaxes the need to compute or use normals; instead, we directly project points to a simple z-buffer. In its basic form, our technique is a *conditional* image-to-image convolutional network. However, a vanilla fully convolutional network would learn translation at the patch level, and ignore long-distance relations. We showed that by injecting positional encoding, we break the locality and gain non-local shading effects.

Limitations and future directions. Our visualizations may not always be completely accurate, for example, nearby points in Euclidean space may be visualized as touching, when in fact they are far in geodesic proximity (see Figure 18). Another problem may arise at fine details, as convolutional neural networks tend to output smooth results. The smoothing effect is also encouraged by the MSE loss our network is trained with, as this loss aims at achieving the *average* visualization result. This problem shows up in Figure 17, where small holes in the cat’s eyes are visualized as closed off. A dedicated experiment for the robustness to small details, using spherical harmonics with increasing frequency, is shown in the in the supplementary material

Our framework offers limited control compared to a full rendering engine. While this can partially be addressed by incorporation textures, materials, and other types of lighting models in the learning procedure, our method should not be seen as a replacement for a rendering system in scenarios where physically-based rendering is more important compared to instant visualization.



Figure 18: Limitation. Our network may have difficulty distinguishing between nearby surfaces, and incorrectly visualization them as a continuous surface. This problem arises from the ambiguity between typical sampling space and surface holes.

There is an interesting trade-off between speed and shading locality that we plan to explore more in the future. Local shading effects, i.e., ambient, diffuse and specular can be learned at the patch level, while global shading, like shadows or reflections, are significantly harder and require learning non-local interactions. One research avenue is to incorporate transformers for learning inter-patches dependencies. Another challenging direction is to improve

the visualization of fine details including sharp features. These, however, are still challenging also for slower methods that reconstruct surfaces from sparse point clouds, let alone for an instant preview.

References

- [ADMG18] ACHLIOPTAS, PANOS, DIAMANTI, OLGA, MITLIAGKAS, IOANNIS, and GUIBAS, LEONIDAS. “Learning representations and generative models for 3d point clouds”. *International conference on machine learning*. PMLR. 2018, 40–49 1.
- [AUL19] ALIEV, KARA-ALI, ULYANOV, DMITRY, and LEMPITSKY, VICTOR. “Neural point-based graphics”. *arXiv preprint arXiv:1906.08240* 2.3 (2019), 4 3, 4.
- [BK03] BOTSCH, MARIO and KOBELT, LEIF. “High-quality point-based rendering on modern GPUs”. *11th Pacific Conference on Computer Graphics and Applications, 2003. Proceedings. IEEE*. 2003, 335–343 1.
- [BMR*99] BERNARDINI, FAUSTO, MITTLEMAN, JOSHUA, RUSHMEIER, HOLLY, et al. “The ball-pivoting algorithm for surface reconstruction”. *IEEE transactions on visualization and computer graphics* 5.4 (1999), 349–359 2, 3.
- [BTS*17] BERGER, MATTHEW, TAGLIASACCHI, ANDREA, SEVERSKY, LEE M, et al. “A survey of surface reconstruction from point clouds”. *Computer Graphics Forum*. Vol. 36. 1. Wiley Online Library. 2017, 301–329 2.
- [CCC*08] CIGNONI, PAOLO, CALLIERI, MARCO, CORSINI, MASSIMILIANO, et al. “MeshLab: an Open-Source Mesh Processing Tool”. *Eurographics Italian Chapter Conference*. Ed. by SCARANO, VITTORIO, CHIARA, ROSARIO DE, and ERRA, UGO. The Eurographics Association, 2008. ISBN: 978-3-905673-68-5. DOI: [10 . 2312 / LocalChapterEvents/ItalChap/ItalianChapConf2008/129-136](https://doi.org/10.2312/LocalChapterEvents/ItalChap/ItalianChapConf2008/129-136) 8.
- [CCCM21] CHEN, XUELIN, COHEN-OR, DANIEL, CHEN, BAOQUAN, and MITRA, NILOY J. “Towards a Neural Graphics Pipeline for Controllable Image Generation”. *Computer Graphics Forum* 40.2 (2021) 3.
- [CGL*19] CHEN, WENZHENG, GAO, JUN, LING, HUAN, et al. “Learning to Predict 3D Objects with an Interpolation-based Differentiable Renderer”. *Advances In Neural Information Processing Systems*. 2019 3.
- [Com21] COMMUNITY, BLENDER ONLINE. *Blender - a 3D modelling and rendering package*. Blender Foundation, Amsterdam, 2021. URL: <http://www.blender.org> 4.
- [CYA*20] CAI, RUOJIN, YANG, GUANDAO, AVERBUCH-ELOR, HADAR, et al. “Learning Gradient Fields for Shape Generation”. *Proceedings of the European Conference on Computer Vision (ECCV)*. 2020 9.
- [DB07] DIANKOV, ROSEN and BAJCSY, RUZENA. “Real-time adaptive point splatting for noisy point clouds.” *GRAPP (GM/R)* 7 (2007), 228–234 2, 4.
- [DZL*20] DAI, PENG, ZHANG, YINDA, LI, ZHUWEN, et al. “Neural point cloud rendering via multi-plane projection”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, 7830–7839 3, 4.
- [EM94] EDELSBRUNNER, HERBERT and MÜCKE, ERNST P. “Three-dimensional alpha shapes”. *ACM Transactions on Graphics (TOG)* 13.1 (1994), 43–72 2.
- [GBZC20] GAL, RINON, BERMANO, AMIT, ZHANG, HAO, and COHEN-OR, DANIEL. “MRGAN: Multi-Rooted 3D Shape Generation with Unsupervised Part Disentanglement”. *arXiv preprint arXiv:2007.12944* (2020) 1.
- [GKOM18] GUERRERO, PAUL, KLEIMAN, YANIR, OVSJANIKOV, MAK, and MITRA, NILOY J. “PCPNet: Learning Local Shape Properties from Raw Point Clouds”. *Computer Graphics Forum* 37.2 (2018), 75–85 2.

- [HB17] HUANG, XUN and BELONGIE, SERGE. “Arbitrary style transfer in real-time with adaptive instance normalization”. *Proceedings of the IEEE International Conference on Computer Vision*. 2017, 1501–1510 3.
- [HDD*92] HOPPE, HUGUES, DE ROSE, TONY, DUCHAMP, TOM, et al. “Surface reconstruction from unorganized points”. *Proceedings of the 19th annual conference on Computer graphics and interactive techniques*. 1992, 71–78 2.
- [HHGC20] HERTZ, AMIR, HANOCKA, RANA, GIRYES, RAJA, and COHEN-OR, DANIEL. “PointGMM: a Neural GMM Network for Point Clouds”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, 12054–12063 1.
- [HMGC20] HANOCKA, RANA, METZGER, GAL, GIRYES, RAJA, and COHEN-OR, DANIEL. “Point2Mesh: A Self-Prior for Deformable Meshes”. *ACM Trans. Graph.* 39.4 (July 2020). ISSN: 0730-0301. DOI: [10.1145/3386569.3392415](https://doi.org/10.1145/3386569.3392415). URL: <https://doi.org/10.1145/3386569.3392415> 2, 3, 7, 8.
- [HMRR19] HERMOSILLA, PEDRO, MAISCH, SEBASTIAN, RITSCHL, TOBIAS, and ROPINSKI, TIMO. “Deep-learning the Latent Space of Light Transport”. *Computer Graphics Forum*. Vol. 38. 4. Wiley Online Library. 2019, 207–217 3.
- [JBG19] JAKOB, JOHANNES, BUCHENAU, CHRISTOPH, and GUTHE, MICHAEL. “Parallel globally consistent normal orientation of raw unorganized point clouds”. *Computer Graphics Forum*. Vol. 38. 5. Wiley Online Library. 2019, 163–173 3.
- [Kaj86] KAJIYA, JAMES T. “The rendering equation”. *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*. 1986, 143–150 3.
- [KBH06] KAZHDAN, MICHAEL, BOLITHO, MATTHEW, and HOPPE, HUGUES. “Poisson surface reconstruction”. *Proceedings of the fourth Eurographics symposium on Geometry processing*. Vol. 7. 2006 2, 3, 7.
- [KH13] KAZHDAN, MICHAEL and HOPPE, HUGUES. “Screened poisson surface reconstruction”. *ACM Transactions on Graphics (ToG)* 32.3 (2013), 1–13 2, 3.
- [KLA19] KARRAS, TERO, LAINE, SAMULI, and AILA, TIMO. “A style-based generator architecture for generative adversarial networks”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, 4401–4410 5.
- [KSL20] KOLOS, MARIA, SEVASTOPOLSKY, ARTEM, and LEMPITSKY, VICTOR. “TRANSPR: Transparency Ray-Accumulating Neural 3D Scene Point Renderer”. *arXiv preprint arXiv:2009.02819* (2020) 3, 4.
- [KTB07] KATZ, SAGI, TAL, AYLLET, and BASRI, RONEN. “Direct visibility of point sets”. *ACM SIGGRAPH 2007 papers*. 2007, 24–es 2, 3.
- [LW10] LIU, SHENGJUN and WANG, CHARLIE CL. “Orienting unorganized points for surface reconstruction”. *Computers & Graphics* 34.3 (2010), 209–218 2.
- [LZZ*18] LI, CHUN-LIANG, ZAHEER, MANZIL, ZHANG, YANG, et al. “Point cloud gan”. *arXiv preprint arXiv:1810.05795* (2018) 1.
- [MDD*10] MULLEN, PATRICK, DE GOES, FERNANDO, DESBRUN, MATHIEU, et al. “Signing the unsigned: Robust surface reconstruction from raw pointsets”. *Computer Graphics Forum*. Vol. 29. 5. Wiley Online Library. 2010, 1733–1741 3.
- [MHGC20] METZGER, GAL, HANOCKA, RANA, GIRYES, RAJA, and COHEN-OR, DANIEL. “Self-Sampling for Neural Point Cloud Consolidation”. *arXiv preprint arXiv:2008.06471* (2020) 1.
- [MHZ*21] METZGER, GAL, HANOCKA, RANA, ZORIN, DENIS, et al. “Orienting Point Clouds with Dipole Propagation”. (2021) 1, 7.
- [MN03] MITRA, NILOY J and NGUYEN, AN. “Estimating surface normals in noisy point cloud data”. *Proceedings of the nineteenth annual symposium on Computational geometry*. 2003, 322–328 2.
- [NAM*17] NALBACH, OLIVER, ARABADZHIYSKA, ELENA, MEHTA, DUSHYANT, et al. “Deep shading: convolutional neural networks for screen space shading”. *Computer graphics forum*. Vol. 36. 4. Wiley Online Library. 2017, 65–78 3.
- [NLBY18] NGUYEN-PHUOC, THU, LI, CHUAN, BALABAN, STEPHEN, and YANG, YONG-LIANG. “Rendernet: A deep convolutional network for differentiable rendering from 3d shapes”. *arXiv preprint arXiv:1806.06575* (2018) 3.
- [PGA11] PINTUS, RUGGERO, GOBBETTI, ENRICO, and AGUS, MARCO. “Real-time rendering of massive unstructured raw point clouds using screen-space operators”. *Proceedings of the 12th International conference on Virtual Reality, Archaeology and Cultural Heritage*. 2011, 105–112 2, 4.
- [PJW12] PREINER, REINHOLD, JESCHKE, STEFAN, and WIMMER, MICHAEL. “Auto Splats: Dynamic Point Cloud Visualization on the GPU.” *EGPGV@ Eurographics*. 2012, 139–148 2, 4.
- [RFB15] RONNEBERGER, OLAF, FISCHER, PHILIPP, and BROX, THOMAS. “U-net: Convolutional networks for biomedical image segmentation”. *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, 234–241 3, 5.
- [RLG*20] RAKOTOSAONA, MARIE-JULIE, LA BARBERA, VITTORIO, GUERRERO, PAUL, et al. “Pointcleannet: Learning to denoise and remove outliers from dense point clouds”. *Computer Graphics Forum*. Vol. 39. 1. Wiley Online Library. 2020, 185–203 1.
- [SLS*06] SHARF, ANDREI, LEWINER, THOMAS, SHAMIR, ARIEL, et al. “Competing fronts for coarse-to-fine surface reconstruction”. *Computer Graphics Forum*. Vol. 25. 3. Wiley Online Library. 2006, 389–398 2.
- [SSG17] SCHERTLER, NICO, SAVCHYNSKY, BOGDAN, and GUMHOLD, STEFAN. “Towards globally optimal normal orientations for large point clouds”. *Computer Graphics Forum*. Vol. 36. 1. Wiley Online Library. 2017, 197–208 3.
- [ST90] SAITO, TAKAFUMI and TAKAHASHI, TOKIICHIRO. “Comprehensible rendering of 3-D shapes”. *Proceedings of the 17th annual conference on Computer graphics and interactive techniques*. 1990, 197–206 3.
- [TSM*20] TANCIK, MATTHEW, SRINIVASAN, PRATUL P., MILDENHALL, BEN, et al. “Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains”. *NeurIPS* (2020) 6.
- [XWZ*21] XIE, CHULIN, WANG, CHUXIN, ZHANG, BO, et al. “Style-based Point Generator with Adversarial Rendering for Point Cloud Completion”. *arXiv preprint arXiv:2103.02535* (2021) 3.
- [YHH*19] YANG, GUANDAO, HUANG, XUN, HAO, ZEKUN, et al. “Pointflow: 3d point cloud generation with continuous normalizing flows”. *Proceedings of the IEEE International Conference on Computer Vision*. 2019, 4541–4550 1.
- [YKH*18] YUAN, WENTAO, KHOT, TEJAS, HELD, DAVID, et al. “Pcn: Point completion network”. *2018 International Conference on 3D Vision (3DV)*. IEEE. 2018, 728–737 1.
- [YLF*18a] YU, LEQUAN, LI, XIANZHI, FU, CHI-WING, et al. “Ec-net: an edge-aware point set consolidation network”. *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, 386–402 1.
- [YLF*18b] YU, LEQUAN, LI, XIANZHI, FU, CHI-WING, et al. “Pu-net: Point cloud upsampling network”. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, 2790–2799 1.
- [YSW*19] YIFAN, WANG, SERENA, FELICE, WU, SHIHAO, et al. “Differentiable surface splatting for point-based geometry processing”. *ACM Transactions on Graphics (TOG)* 38.6 (2019), 1–14 3, 8.
- [YWH*19] YIFAN, WANG, WU, SHIHAO, HUANG, HUI, et al. “Patch-based progressive 3d point set upsampling”. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, 5958–5967 1.
- [ZPVG01] ZWICKER, MATTHIAS, PFISTER, HANSPETER, VAN BAAR, JEROEN, and GROSS, MARKUS. “Surface splatting”. *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. 2001, 371–378 1, 2, 8.