

# Interactive Annotation of 3D Object Geometry using 2D Scribbles

Tianchang Shen<sup>1,2\*</sup>, Jun Gao<sup>1,2,3\*</sup>, Amlan Kar<sup>1,2,3</sup>, and Sanja Fidler<sup>1,2,3</sup>

<sup>1</sup>University of Toronto, <sup>2</sup>Vector Institute, <sup>3</sup>Nvidia  
{shenti11, jungao, amlan, fidler}@cs.toronto.edu

**Abstract.** Inferring detailed 3D geometry of the scene is crucial for robotics applications, simulation, and 3D content creation. However, such information is hard to obtain, and thus very few datasets support it. In this paper, we propose an interactive framework for annotating 3D object geometry from both point cloud data and RGB imagery. The key idea behind our approach is to exploit strong priors that humans have about the 3D world in order to interactively annotate complete 3D shapes. Our framework targets naive users without artistic or graphics expertise. We introduce two simple-to-use interaction modules. First, we make an automatic guess of the 3D shape and allow the user to provide feedback about large errors by drawing scribbles in desired 2D views. Next, we aim to correct minor errors, in which users drag and drop mesh vertices, assisted by a neural interactive module implemented as a Graph Convolutional Network. Experimentally, we show that only a few user interactions are needed to produce good quality 3D shapes on popular benchmarks such as ShapeNet, Pix3D and ScanNet. We implement our framework as a web service and conduct a user study, where we show that user annotated data using our method effectively facilitates real-world learning tasks. Web service: <http://www.cs.toronto.edu/~shenti11/scribble3d>.

## 1 Introduction

3D scene understanding is a crucial component of numerous robotic applications such as autonomous driving, household robots, and delivery drones. In order to successfully plan its next move, an agent needs to infer 3D geometry of both the scene and relevant objects. Furthermore, the inferred 3D geometry should be sufficiently detailed to afford fine-grained interaction such as manipulation and grasping. We stress that reasoning about the visible scene alone is insufficient: objects occupy physical space in the world, extending beyond what is visible, which needs to be taken into account for downstream tasks. Reconstructing detailed and complete geometry of each object is also important for simulation, where the goal is to convert scanned point cloud scenes into interactive virtual environments to train artificial agents in, prior to real world deployment.

We need datasets in order to infer complete 3D object geometry from partial observations (point clouds or imagery), which have been collected with two

---

\* equal contribution

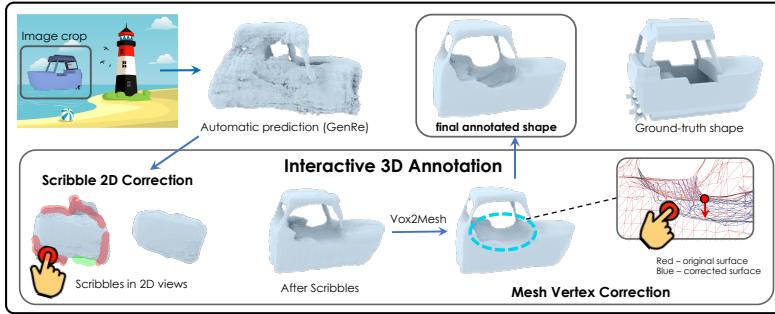


Fig. 1: Interactive 3D Annotation Framework: Our framework has two interaction modules: **1) large errors**: user provides feedback by drawing scribbles in desired 2D views – we exploit this to predict a better shape, **2) small errors**: user drags & drops mesh vertices, and we repredict the mesh locally. Invisible object regions are up to user’s priors/imagination.

dominant approaches. In [6], the authors exhaustively scan each individual object, followed by a point-registration method. However, detailed scanning is time consuming and limited to static scenes and objects that can be scanned from all possible views – which may require lifting the object from its support surface. The second approach employs fitting CAD models to segmented point clouds [8, 16, 17] or RGB images [21, 32]. This is an undesirable alternative as it requires a large and diverse collection of CAD models to faithfully match input shapes. Such collections are not always available for any object class.

In this paper, we propose an interactive framework for annotating 3D object geometry from both point clouds and RGB images (Fig. 1). At its core, we exploit the extensive prior knowledge in humans to complete (annotate) objects. Since 3D CAD model creation typically requires artistic expertise, our key contribution lies in an interactive learning-based approach that enables a graphics layman to produce good quality 3D meshes that fit (partial) 3D observations.

Our annotation framework is composed of two stages. First, we employ existing methods to produce an initial guess of the 3D shape. To leverage the fact that 2D interactions are easier on existing digital media than 3D, we introduce a simple interface to correct large 3D errors by scribbling corrections in one or several desired 2D views. We propose a neural architecture that uses these scribbles to produce an improved prediction of the 3D shape. Next, for final fine-grained corrections, the annotator makes edits directly to 3D vertices of the object mesh using a drag-and-drop interface. Our 3D Graph Convolutional Network takes this correction into account and aims to re-predict refined local geometry, while ensuring that the majority of the object geometry remains intact.

We demonstrate that our interactive approach, **1)** is able to generate high-quality meshes from single images on the ShapeNet [3] and Pix3D [29] datasets, even on

unseen object categories with very few interactions from an user. The annotated shapes can facilitates real-world learning tasks, **2)** comparing favourably to existing annotation strategies such as those employing CAD model retrieval and **3)** can be used for full-scene annotation on the popular ScanNet dataset [6].

## 2 Related Work

**3D Reconstruction:** The problem of recovering 3D shape from RGB-D or RGB images is challenging due to the inherent ambiguity in going from a lower to a higher dimensional space. Existing learning-based works can be categorized into four different categories based on output representation [19]: voxel [7, 28], point cloud [10], mesh [5, 14, 30] and signed distance field [25, 34]. In practice, by design, these approaches memorize the classes seen during training and exhibit limited generalization ability to unseen classes. Recently, GenRe [35] proposed to exploit spherical maps in order to perform depth completion which has been shown to achieve better generalization to unseen classes. Our framework employs GenRe for its generalization ability, but is not limited to it, *i.e.* other methods can be used. Our key contribution is in making these methods interactive.

**Interactive 3D Modeling** is a long studied topic in computer graphics [4, 9, 13, 27]. Image-based methods assist the user to snap parametric primitives (*e.g.* generalized cylinders, ellipsoids) to image contours [4, 12, 27]. A combination of such primitives forms the final shape. [13] further allows users to indicate higher-level semantic information, *e.g.* equal length/angles, symmetries. The major drawback of these methods is that they require the user to mentally imagine how to decompose 3D shape into primitives and perform their composition. This is not an easy task for a naive user. In our paper, we propose a learning-based interactive approach that assists a graphics layman to annotate complex shapes.

**Data-driven approaches** have also been proposed to create 3D models. Early works mainly perform retrieval from a database, by either retrieving a whole 3D shape [11], or procedurally obtain predefined assemblies [20, 33]. [23] proposed to learn the correlation between 2D sketches and 3D shape. Recently, [9] utilized a deep network to learn this correlation, and users are able to modify the 3D shape via 2D sketches. However, drawing detailed sketches requires some level of artistic expertise. [24] extend the idea of [36] into 3D by utilizing a GAN to learn a latent representation of 3D shapes. User’s edits are mapped to this latent space and the network regenerates a shape. This approach loses control over the generation and the network easily neglects user’s corrections.

## 3 Interactive 3D Annotation

We now describe our framework for annotating 3D object. Key desired properties include: **1)** naturally incorporate user interactions, **2)** generalize to unseen object classes and **3)** produce high-quality shapes. We emphasize that generalization beyond training shapes is crucial to scale to the diversity in the real-world. Since

our goal is to produce annotated data to power applications required to predict high quality 3D shapes (*e.g.* VR/AR, simulation, manipulation in robotics), we want our framework to support annotation of detailed geometry. Furthermore, we want our interactive tool to be seamlessly usable by anyone, without assuming expertise in 3D modeling, which is key in undertaking large-scale annotation endeavours in the future.

Our method supports annotating 3D geometry from RGB images and point clouds. It is designed to reconstruct a complete 3D object shape in a coarse-to-fine manner, incorporating annotator corrections at all stages. In particular, it consists of two modules. The *Scribble Interaction Module* (SIM), allows the user to make dramatic changes to the annotated shape by drawing coarse scribbles. The user inspects the 3D shape by rotating it to any viewpoint, and makes corrections to the projected silhouette by drawing scribbles in 2D, which is used to re-predict the shape. The user iterates until the desired quality is achieved. Next, the *Point Interaction Module* (PIM) allows the user to make fine corrections directly in 3D by moving vertices of the object’s mesh to obtain fine-grained geometric details. We use this feedback to re-predict the local mesh geometry in order to minimize human effort in editing the final mesh.

In Section 3.1, we describe our annotation setup, including the initial automatic prediction of the 3D shape. In Section 3.2, we introduce our Scribble Interaction Module. Finally, we describe our Point Interaction Module in Section 3.3.

### 3.1 Annotation Setup

We envision a human user annotating 3D shapes for possibly multiple objects in an image (or a point cloud scene). The user is expected to indicate which object to annotate by drawing a 2D mask overlaying the object. Note that this can be done very quickly by using existing interactive techniques [22, 31]. Given the selected object, we aim to predict and assist in the interactive annotation of an accurate 3D shape. Our framework can incorporate any existing 3D prediction network, and our main contribution is in making annotation interactive.

**Automatic 3D Shape Prediction:** We choose GenRe [35] as our automatic prediction network, for its ability to generalize to beyond training shapes. We briefly summarize it here and refer readers to the original paper for details. GenRe uses three steps: 1) predicting a depth map from an RGB image (cropped to contain a single object), 2) converting the depth map into a spherical map and inpainting the missing depth information with a 2D CNN and 3) refining the 3D shape. The first module is optional and is used when the input is a monocular image. With access to a depth map or a point cloud, this step can be omitted. In the third module, the completed spherical depth map from the second module is back-projected into a 3D occupancy grid, which is further refined to a volume of size  $128^3$  using a 3D-UResNet (visualization is provided in the Appendix).



We note that the resulting shape is viewer-centric, allowing us to easily overlay it onto sensor inputs (*e.g.* RGB image, point cloud), which is useful for the user tasked to provide corrections to the predicted shape. When depth is available, we can place the object in camera coordinates, *i.e.* into the world scene.

### 3.2 Scribble Interaction Module

The Scribble Interaction Module (SIM) helps correct major errors in the initial 3D shape prediction. It is purposefully designed to be 2D view-based, targeting an average 3D illiterate user. Unless specially trained to craft 3D content such as artists or graphics experts, a naive user is known to be better at 2D editing. SIM then learns to propagate 2D corrections to 3D in order to refine the shape.

**Annotation Setup:** To make edits to a shape (automatic prediction initially), the user views it in 3D and rotates it to any desired viewpoint. If a view with a major flaw is discovered, the user can indicate errors by drawing scribbles onto the (2D) projection (Figure 2 provides an illustration). Following existing work on scribble-based image annotation [26], we support “additive” scribbles to indicate false negative areas and “deletion” scribbles to indicate false positive areas. A scribble contains pixels obtained by dilating the trace of the mouse cursor with a kernel of a user-specified bandwidth, called the scribble width. When the user wants to add (or remove) fine details to the 3D shape, a small scribble width can be used. Nonetheless, this interaction is coarse in the sense that a scribble typically does not accurately trace high frequency regions.

**Scribble-based Neural Module:** Since the scribble is marked in 2D, there is an inherent ambiguity in 3D associated with 2D correction. We design a neural architecture that makes a best guess of which voxels in the 3D volume should be corrected based on 2D scribbles. The annotation module allows the user to iteratively rotate the refined shapes and indicate errors, until a satisfactory quality is achieved. Our architecture (Figure 3) is based on the intuition that predicting 2D projections of a shape is likely an easier task compared to predicting a 3D occupancy grid. The network needs to infer how the refined silhouette of the shape would look in the view in which the scribbles were drawn, and other related views (*e.g.* two orthogonal views), while exploiting depth information and user scribbles. The shape can then be obtained by 3D carving using the predicted silhouettes, and refining the “carved” shape with a simple architecture.

**Overview of Architecture:** The model takes as input a 3D feature map (from GenRe) and rotates it to align with the user-specified view. We perform max-pooling along the  $z$  axis to obtain a *2D-view feature representation* of the input projected into user’s view. Scribbles are encoded along with the projected feature map to obtain a new representation of the corrected viewpoint using a *2D-scribble encoder*. In order to propagate this information to the orthogonal

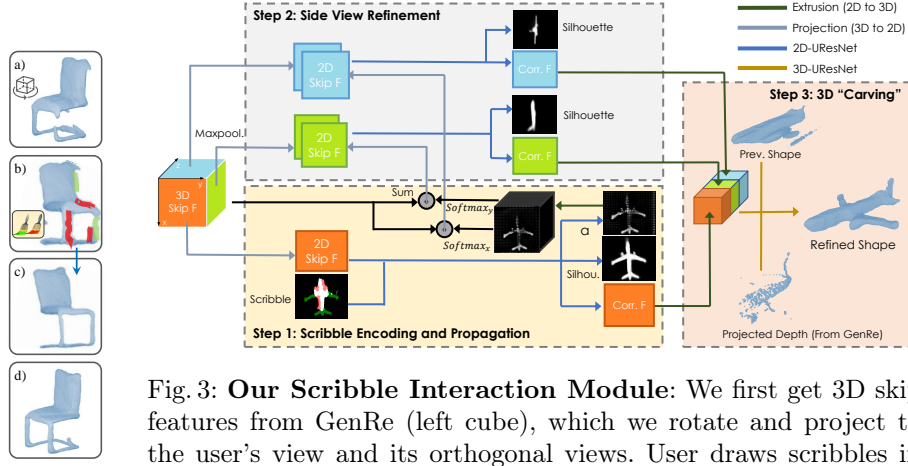


Fig. 2: 2D Scribble Annotation.

Fig. 3: **Our Scribble Interaction Module:** We first get 3D skip features from GenRe (left cube), which we rotate and project to the user’s view and its orthogonal views. User draws scribbles in the 2D view to indicate errors in the predicted shape. Our network encodes and propagates scribbles to orthogonal views. Lastly, we concatenate refined features from all three orthogonal views, along with the shape from the previous step and the projected depth, to predict a new refined shape.

views, we similarly project GenRe’s feature map onto the two other views, and concatenate information propagated from the user-specified view to each of the projected feature maps, using the *scribble propagation module*. We decode these concatenated maps into the final view representations using a 2D-UResNet. To predict the final 3D shape, the *3D carving module* performs an operation mimicking carving in the feature space to obtain a 3D volumetric representation which is decoded into a 3D occupancy grid. We provide details next.

**2D-view Feature Representation:** Let  $R$  denote the rotation that rotates the object into the user’s viewpoint. We extract skip feature maps,  $F \in \mathbb{R}^{C \times 64 \times 64 \times 64}$ , from the first and last 3D convolutional layers of GenRe’s 3D-UResNet and rotate it to the user’s 3D viewpoint via  $R$ . The feature map is upsampled by a factor of 2 using bilinear interpolation, denoted by  $F_v \in \mathbb{R}^{C \times 128 \times 128 \times 128}$ .  $F_v$  is passed through a single FC layer and is projected to three orthogonal views by max-pooling along  $x$ ,  $y$ , and  $z$  axes, *i.e.*  $F_v^{xy} = \text{maxpool}_z(F_v)$  (similar for  $F_v^{yz}$  and  $F_v^{xz}$ ). Pooling along  $z$  mimics an orthographic projection into the user’s view, while pooling along  $x$  and  $y$  projects into its orthogonal views. While more elaborate pooling operations could be performed mimicking perspective projection, we found the simpler approach to work well in practice.

**Iterative Correction:** Note that our interactive approach runs iteratively, allowing the user to draw scribbles in a different viewpoint each time. In the first correction round, we use 3D features from GenRe, while in all subsequent

rounds we use features from SIM. However, since GenRe’s and SIM’s outputs differ, we choose to use 2 different weight matrices for the first vs other steps. The 3D refinement network weights are shared across all correction rounds.

**Scribble Encoder:** We encode positive/negative scribbles by concatenating them along with the initial (binary) silhouette into  $S \in \mathbb{R}^{3 \times 128 \times 128}$ . To process scribbles, we use a 2D-UResNet, that takes a concatenation of the projected 2D feature  $F_v^{xy}$  and  $S$  as input and predicts a refined silhouette  $M^{xy}$  as well as refined features  $F_c^{xy}$  in the user’s view:

$$F_s = \text{concat}\{F_v^{xy}, S\}, \quad (1)$$

$$M^{xy}, F_c^{xy} = \text{2D-UResNet}(F_s) \quad (2)$$

$F_c^{xy}$  is used in the 3D carving module, while the mask prediction is used for an auxilliary loss function during training.

**Scribble Propagation Module** aims to propagate scribble information from the user’s view into orthogonal views. We utilize a 2D-UResNet to predict an attention map  $\mathbf{A}^{xy} = \text{2D-UResNet}(F_s) \in \mathbb{R}^{128 \times 128}$ , which functions as a gate to silence or amplify features of the entire tube. We then extrude the attention map along  $z$ -axis:  $\hat{\mathbf{A}}^{xy} \in \mathbb{R}^{128 \times 128 \times 128}$ , and perform softmax along different axes when propagating to the orthogonal views:

$$\hat{\mathbf{A}}_{i,j,k}^{yz} = \frac{\exp(\hat{\mathbf{A}}_{i,j,k}^{xy})}{\sum_i \exp(\hat{\mathbf{A}}_{i,j,k}^{xy})}, \hat{\mathbf{A}}_{i,j,k}^{xz} = \frac{\exp(\hat{\mathbf{A}}_{i,j,k}^{xy})}{\sum_j \exp(\hat{\mathbf{A}}_{i,j,k}^{xy})} \quad (3)$$

where  $i, j, k$  denote 3D tensor indices. We compute the weighted feature map:

$$\hat{F}_v^{yz} = F_v \circ \hat{\mathbf{A}}^{yz}, \hat{F}_v^{xz} = F_v \circ \hat{\mathbf{A}}^{xz}, \quad (4)$$

where  $\circ$  denotes the Hadamard product.  $M^{yz}, F_c^{yz}, M^{xz}, F_c^{xz}$  are predicted in a similar manner as in the user’s view, but taking projected features from both  $F_v$  and the weighted feature map. For the  $yz$ -plane features (similar for  $xz$ ),

$$M^{yz}, F_c^{yz} = \text{2D-UResNet}(\text{concat}\{\sum_x (\hat{F}_v^{yz}), F_v^{yz}\}) \quad (5)$$

**3D Carving Module:** The refined view features are backprojected to 3D by extruding  $F_c^{xy}, F_c^{yz}$  and  $F_c^{xz}$  along their orthogonal axes. A simple refinement network takes the concatenation of the three extruded feature maps, previous occupancy grid and the projected depth map as input, and outputs a new volume as the final prediction. The projected depth map is included to preserve consistency with sensor observations. Note that we do not explicitly use the predicted silhouettes, but instead use the predicted feature maps. We argue that the feature maps preserve more information and thus are better for 3D refinement, while predicting and supervising silhouettes helps learn better features.

### 3.3 Point Interaction Module

A voxelgrid with limited resolution lacks fine geometric details. We thus convert the predicted occupancy grid to a triangular mesh and allow the annotator to drag-and-drop vertices, one at a time, and employ the point interaction module to locally refine the mesh based on user feedback. To convert the occupancy grid to a mesh, a Vox2Mesh method similar to [14, 30] is used, with minor but important modifications that yield visibly better meshes (details in suppl.).

**Point-based Neural Module** takes the user’s vertex correction and employs a GCN to repredict its neighboring vertices. As shown in the Fig. 4, we use a similar network architecture to [22], which is designed for interactive 2D curve correction. In particular, it performs:

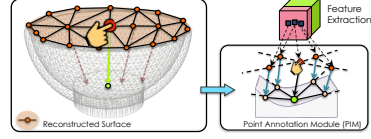


Fig. 4: PIM allows users to interactively edit a mesh.

$$f_i = \text{concat}(F(p_i); p_i; q_i - p_i), \quad (6)$$

$$\Delta p_j = \text{GCN}(f_1, f_2, \dots, f_N; G), \quad (7)$$

where  $F(p_i)$  represents the feature bilinearly sampled from 3D feature map with position  $p_i$ , which is corrected to  $q_i$ ,  $N$  is the number mesh vertices, and  $G$  is the topology of the mesh.  $\Delta p_j$  represents the predicted offset for each local point  $p_j$ . We restrict movement to only neighbours along paths of length  $l$  from the edited vertex. It is trained using similar loss as MeshRCNN [14] (see appendix).

## 4 Experiments

In this section, we provide extensive evaluation of our model. To demonstrate the model’s capability as an annotation tool, we report interactive single image shape reconstruction results on ShapeNet for both seen and unseen classes in Sec. 4.2. We further evaluate our full pipeline on the challenging Pix3D [29], ScanNet [8] and Scan2CAD [2] dataset to show the generalizability of our method to real-world sensor inputs in Sec. 4.3. In addition, we provide quantitative analysis comparing our method to CAD Retrieval using ShapeNet models to annotate Pix3D instances in Sec. 4.3. We show ablation studies comparing SIM with other network architectures in Sec. 4.4. Finally, we annotate real data in a user study and show results of models trained on this data in Sec. 4.5.

### 4.1 Experimental Settings

For all experiments, we train our model on Car, Chair and Airplane, which are the three largest categories in ShapeNet [3]. We use the same subset split and rendered images as in [35]. For evaluation on real-world data, we use the

		Seen Classes	Unseen Classes									
			Bch	Vsl	Rfl	Sfa	Tbl	Phn	Dsp	Spk	Lmp	Avg
Chamfer Dist.	GenRe	0.1093	0.1158	0.1477	0.1570	0.1140	0.1572	0.1522	0.1921	0.1738	0.2667	0.1640
	+ Scribble Annot.	0.0503	0.0435	0.0517	<b>0.0289</b>	0.0465	0.0565	<b>0.0315</b>	<b>0.0461</b>	<b>0.0630</b>	0.1190	<b>0.0541</b>
	+ Vox2Mesh	<b>0.0474</b>	<b>0.0412</b>	<b>0.0474</b>	0.0301	<b>0.0452</b>	<b>0.0545</b>	0.0345	0.0499	0.0681	<b>0.1178</b>	0.0543
F1 Score	GenRe	0.8392	0.8211	0.7270	0.6837	0.8368	0.7221	0.6993	0.6014	0.6832	0.4927	0.6963
	+ Scribble Annot.	0.9743	0.9806	<b>0.9787</b>	0.9916	0.9629	0.9629	0.9871	<b>0.9774</b>	<b>0.9490</b>	0.8307	0.9578
	+ Vox2Mesh	<b>0.9757</b>	<b>0.9849</b>	0.9780	<b>0.9854</b>	<b>0.9858</b>	<b>0.9642</b>	<b>0.9885</b>	0.9714	0.9387	<b>0.8320</b>	<b>0.9587</b>
Normal Consist.	GenRe	0.6353	0.6131	0.5640	0.4945	0.6266	0.6202	0.6029	0.5632	0.6030	0.5004	0.5764
	+ Scribble Annot.	0.7954	0.7633	0.7522	<b>0.7454</b>	0.8429	0.8171	0.8907	0.8506	0.8371	<b>0.7031</b>	0.8002
	+ Vox2Mesh	<b>0.8128</b>	<b>0.7905</b>	<b>0.7563</b>	0.7257	<b>0.8477</b>	<b>0.8360</b>	<b>0.8946</b>	<b>0.8692</b>	<b>0.8386</b>	0.7002	<b>0.8065</b>

Table 1: Reconstruction results of the training classes and 9 unseen classes from ShapeNet. With our interactive method, we consistently improve and significantly improve over the automatic 3D reconstruction approach.

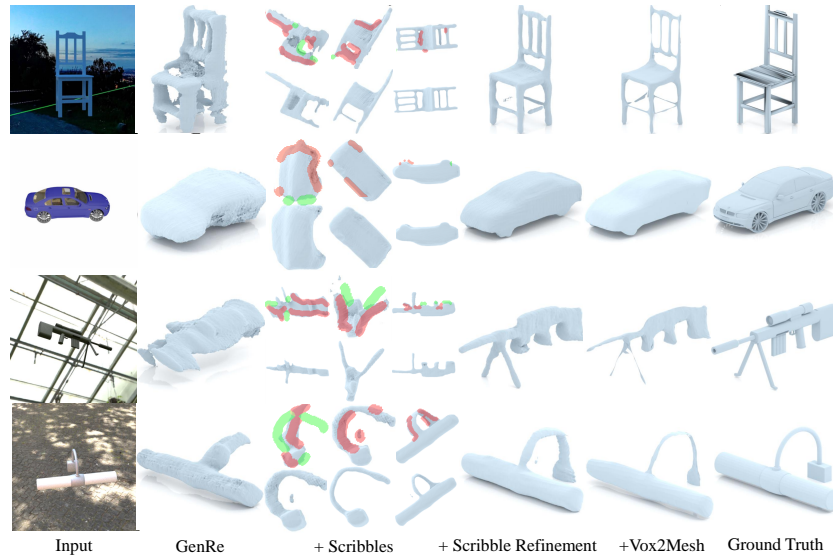


Fig. 5: Examples of ShapeNet objects annotated using our framework. First two rows are from seen classes while others are from unseen classes. Output shape from each intermediate step is shown from left to right, followed by GT mesh. Additive/deletion scribbles are shown in green/red. Although the additive scribbles are drawn in 2D, SIM accurately infers the depth of added shape (e.g. bipods in the third example). With a few scribbles, our model predicts shape close to GT. Vox2Mesh further converts a discrete voxel grid to a fine mesh.

Pix3D [29] and ScanNet [8]. ScanNet is a richly annotated indoor dataset that provides RGB-D sequences, cameras, surface reconstructions, and instance-level semantic segmentation. It further provides CAD models aligned to the objects on Scan2CAD [2]. During inference, we use 5 correction steps for both interaction methods unless otherwise mentioned. More details are in appendix.

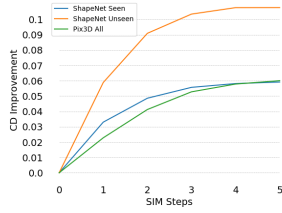


Fig. 6: Improvements in Chamfer Dist vs num of scribble refin. steps.

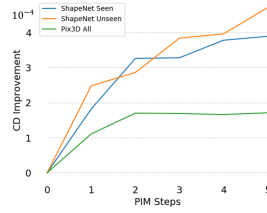


Fig. 7: Improvements in Chamfer Dist vs. # of point annotation steps.

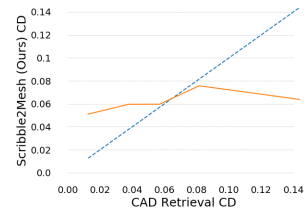


Fig. 8: Chamfer Dist of the best-aligned CAD model vs. result obtained by our method.

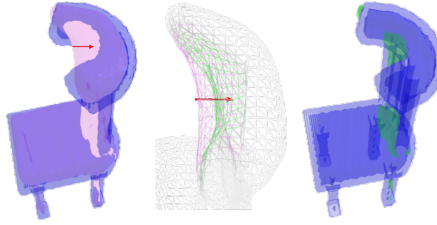


Fig. 9: An example with PIM. The ground truth, initial prediction and refined prediction are colored in blue, pink and green, respectively. The red arrow represents correction for chosen vertex. The entire local surface is pushed towards the ground truth surface, demonstrating that PIM can annotate concave surfaces effectively.

## 4.2 ShapeNet Annotation

**SIM:** To evaluate SIM, we simulate user scribbles as in training (see appendix). We report results on both seen and unseen classes in Tab 1. Improvement in Chamfer Distance as the corrections are sequentially applied is shown in Fig. 6. With only coarse annotation, SIM significantly improves raw predictions under all metrics, and across all object categories with higher improvement in unseen classes. This demonstrates that our method generalizes to novel objects, making it suitable for the annotation task. We provide qualitative examples in Fig. 5, demonstrating that with a few scribbles, our proposed module can significantly improve the initial prediction, bringing it very close to the ground-truth shape.

**PIM:** PIM is used to further refine the mesh (from Vox2Mesh). Following [1,22], we use the same method to simulate the annotator to correct vertices in training. We report improvements in Chamfer Distance in Fig. 7. As a local correction method, applying point annotation gradually improves the quality of 3D shape. Qualitative result in Fig. 9 shows that with 1 correction, PIM can effectively push the prediction to a concave surface, which cannot be annotated by scribbles.

		Chair	Bed	Bookcase	Desk	Sofa	Table	Wardrobe
Chamfer Distance	GenRe	0.1161	0.1427	0.1167	0.1526	0.0939	0.1735	0.1128
	+ Scribble Annotation	0.0609	<b>0.0696</b>	<b>0.0546</b>	0.0748	0.0596	0.0968	<b>0.0413</b>
	+ Vox2Mesh	<b>0.0588</b>	0.0703	0.0575	<b>0.0705</b>	<b>0.0579</b>	<b>0.0915</b>	0.0458
F1 Score	GenRe	0.8251	0.7483	0.8126	0.7145	0.8957	0.6667	0.8302
	+ Scribble Annotation	0.9549	<b>0.9362</b>	<b>0.9660</b>	0.9306	0.9720	0.8775	<b>0.9851</b>
	+ Vox2Mesh	<b>0.9591</b>	0.9358	0.9614	<b>0.9345</b>	<b>0.9779</b>	<b>0.8808</b>	0.9819
Normal Consistency	GenRe	0.6217	0.6479	0.5673	0.6110	0.6541	0.6347	0.6462
	+ Scribble Annotation	0.7358	0.7935	0.7022	0.7807	<b>0.8160</b>	0.7784	<b>0.8776</b>
	+ Vox2Mesh	<b>0.7587</b>	<b>0.8000</b>	<b>0.7037</b>	<b>0.7990</b>	0.8147	<b>0.7969</b>	0.8628

Table 2: Reconstruction results on the Pix3D Objects dataset.

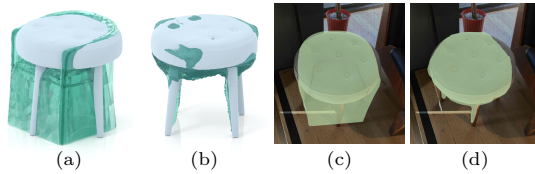


Fig. 10: Example of CAD alignment (a) vs. our approach (b). Light blue shape is the ground truth. We render both shapes on the input image at c) and d). Our method is more consistent with both ground-truth shape and input.

### 4.3 Annotating Real Scans

**Pix3D Annotation:** We now evaluate our trained model on Pix3D [29]. We do not fine-tune the model, and only run inference on this dataset. We test on unoccluded and un-truncated shapes and use ground-truth 3D shape to simulate scribbles. Quantitative results are shown in Tab. 2. Our model achieves the same level of improvement across all Pix3D categories despite the domain gap between synthetic images rendered from ShapeNet models and real imagery.

We compare with a common alternative approach for 3D annotation of retrieving and aligning CAD models from a large database [2, 29]. For fair comparison, we use our training and validation sets as the retrieval database. For retrieval, we pick the object with the minimum Chamfer Distance to the ground-truth CAD model (or scanned surface). Comparison on all 200 chair models in Pix3D is shown in Fig. 8. CAD retrieval is promising if the database contains similar objects, but it fails for unseen objects, which is crucial in practice. Our model overcomes this limitation and performs consistently regardless of the similarity to the training objects (Fig. 10). Our method achieves more faithful reconstruction to unseen objects, and can be used alongside CAD retrieval in a real application.

### 4.4 Analysis

**Ablation Study:** We explore two alternatives for scribble propagation. We could directly extrude 2D scribbles to 3D, or learn to encode the 2D scribbles in the correction view and extrude the learned feature directly to 3D (avoiding prediction of 2D side views in our model). The second approach is a generalization



of the updater network in [9], combined with volumetric aggregation in [18] to avoid repeated rotation of the feature space. We train these with the same training scheme and 3D U-ResNet architecture. The result is summarized in Table 3. Our proposed multi-view correction module outperforms both alternatives in all metrics, and the reconstruction is more plausible and complete (see Fig. 11).



Fig. 11: (From left to right) The same chair annotated by 3D-scribble, single-view and multi-view SIM respectively.

	Seen			Unseen		
	CD	F1	Normal	CD	F1	Normal
3D-Scribble	0.0628	0.9539	0.7729	0.0804	0.9109	0.7574
1-View	0.0555	0.9652	0.7795	0.0609	0.9499	0.7791
Multi-View	<b>0.0535</b>	<b>0.9694</b>	<b>0.7899</b>	<b>0.0589</b>	<b>0.9538</b>	<b>0.7838</b>

Table 3: Comparison of different SIM architectures.

**Shape Creation with Scribbles:** Our model is capable of creating a plausible shape purely from scribble inputs, as shown in Fig. 12. This mimics the scenario where the user wants to create 3D shapes by simply drawing scribbles on a blank screen. As before, we allow the user to iteratively add scribbles from multiple views in order to obtain the desired shape. In the example, the user is able to create a plausible full chair with only 3 drawing steps.

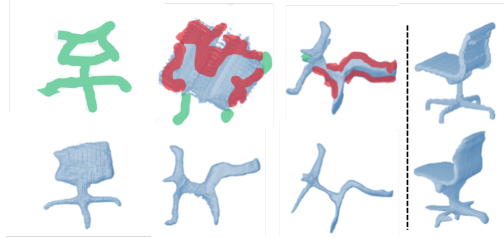


Fig. 12: Shape creation from only scribble inputs. (top) Images on the left show scribbles, (bottom) object in correction view after refinement. The ground truth is shown on top right.

#### 4.5 User Study

To evaluate our method with real human annotators, we built a web tool visualized in Fig. 13, and conducted a user study on two real-world datasets: Pix3D [29] and ScanNet [6]. The annotator engaged in the user study had no prior experience in 3D modeling and did not have access to ground truth shapes. The tool only uses scribble corrections with SIM. Details can be found in appendix.



Fig. 13: Screenshot.

**Full Scene Annotation:** We show full scene annotation capability of our approach on ScanNet [6]. We compare with CAD

alignment in Fig. 14 (Scan2CAD provides human-annotated CAD alignment as ground truth). Annotators took 2.48 minutes on avg. to find the appropriate CAD model and align it in the scene. We report an avg. time of 1.88 minutes in Tab. 4. Our model faithfully reconstructs the surface of all objects in the scene, aligning better with the partial surface from the scan than CAD retrieval.

	Bed	Bookcase	Desk	Sofa	Table	Wardrobe	Avg.
Avg. # Scrib.	15	18	10	10	19	9	14
Avg. Time(sec)	141	109	116	71	184	55	113
CD Before	0.1555	0.1048	0.1330	0.0893	0.1637	0.1043	0.1251
CD After	0.0859	0.0517	0.0778	0.0746	0.0989	0.0513	0.0733

Table 4: Statistics on Pix3D.

	Bed	Bookcase	Desk	Sofa	Table	Wardrobe	Avg.
Pretrained	0.1728	0.1603	0.2581	0.0998	0.1659	0.1863	0.1738
Tuning on Ours	0.1094	0.1017	0.1642	0.0870	0.1279	0.0898	0.1133
Tuning on GT	0.0712	0.0776	0.1168	0.0800	0.1023	0.0555	0.0839

Table 5: Reconstruction results of OccNet on Pix3D before and after fine-tuning.

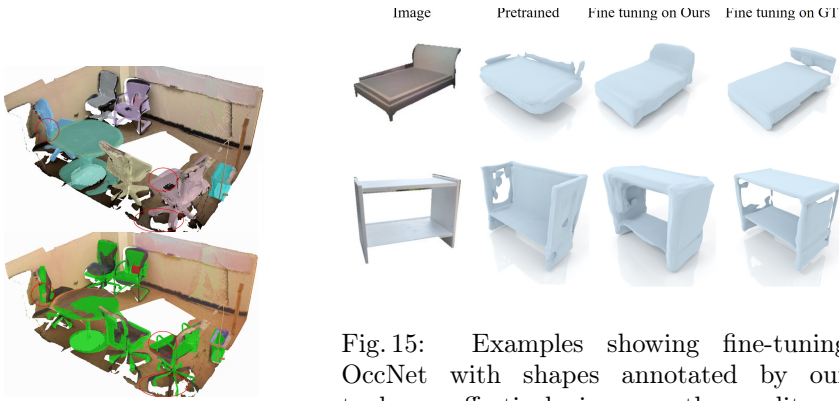


Fig. 14: Full scene annotation (top) vs. CAD alignment (bottom) on ScanNet.

Fig. 15: Examples showing fine-tuning OccNet with shapes annotated by our tool can effectively improve the quality of reconstruction on target task. Note that the *bed* category (top) is unseen in the pretrained OccNet model.

**Real Dataset Creation:** To validate if our annotated data is useful for learning, we create a dataset from Pix3D images and use it to fine-tune a 3D reconstruction model pretrained on synthetic data. In this study, we chose to annotate all unseen categories in Pix3D, which has 2817 images containing un-occluded and un-truncated objects, corresponding to 147 different CAD shapes. The annotator was asked to annotate 95 randomly-selected shapes from random views and the remaining shapes were used for testing. Annotation results are shown in Tab. 4. We found that the large errors in the initial prediction were effectively fixed by user corrections, but the surface is often jagged due to human error. Qualitative examples and screen recordings of the annotation process are available in the appendix. Note that runtimes of SIM and Vox2Mesh are 0.49s and 0.007s, respectively. The majority of the time is spent on inspecting shapes and drawing scribbles. The improvement on Chamfer Distance is comparable to that using simulated scribbles, showing that our simulation is representative of real human inputs. In addition, we found humans can infer invisible (occluded) 3D geometry from a single image much better than the SOTA 3D reconstruction models, which is the gap that we aim to bridge with our proposed interactive annotation tool.

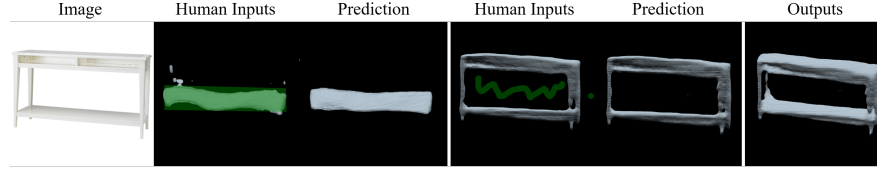


Fig. 16: (left): Example of user annotating with a polygon, which traces the boundary instead of the skeleton of the negative region. (right): Example when the human provides wrong scribbles. Notice that this scribble is being largely ignored since the network tries to preserve the integrity of the shape.

We used the 95 annotated shapes with 1880 corresponding images to fine-tune OccNet [25] (instead of GenRe, to avoid model bias) pre-trained on ShapeNet, and report performance in Tbl. 5 with qualitative examples in Fig. 15. Despite the remaining artifacts due to annotation errors, the integrity of the reconstructed shape is largely improved, which demonstrates that our annotation tool is a step towards facilitating real-world 3D annotation for learning. Note that although SIM is trained with simulated scribbles, our model shows **robustness to human input** *i.e.* to human annotation errors or noise (see Fig. 16).

**Limitation and Future Work:** During our user study, we found it was difficult to annotate 1) complex shapes from bad initial predictions and 2) small but fine details. The *addition* signal from the user is sometimes falsely ignored, causing repeated scribbles at the same location. Vox2Mesh sometimes creates artifacts along thin edges (*e.g.*, chair leg, airfoil), which is commonly observed in deformation-based mesh generation methods [14, 15, 30]. We include examples of these failure cases in the appendix. To further improve our model, a possible direction is to apply the interaction in other form of representations, such as implicit function, instead of volumetric space to improve the quality of the generated shape, and avoid the additional step of converting voxels to a mesh.

## 5 Conclusion

In this paper, we provide a 3D annotation framework with two simple-to-use interaction methods, where user can correct 3D shape with scribbles and drag-and-drop clicks, and our model takes these feedback and refines 3D prediction. Experiments on both synthetic images and real sensory data demonstrate the effectiveness of our method. Our tool is available as a webservice to facilitate the progress of 3D dataset collection for 3D reasoning.

**Acknowledgments.** We thank Louis Clergue for assistance with developing the web tool and extended discussion. This work was supported by NSERC. SF acknowledges the Canada CIFAR AI Chair award at the Vector Institute.

## References

1. Acuna, D., Ling, H., Kar, A., Fidler, S.: Efficient interactive annotation of segmentation datasets with polygon-rnn++. In: CVPR. pp. 859–868 (2018) [10](#)
2. Avetisyan, A., Dahnert, M., Dai, A., Savva, M., Chang, A.X., Niessner, M.: Scan2cad: Learning cad model alignment in rgb-d scans. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) [8](#), [9](#), [11](#)
3. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015) [2](#), [8](#)
4. Chen, T., Zhu, Z., Shamir, A., Hu, S.M., Cohen-Or, D.: 3-sweep: Extracting editable objects from a single photo. ACM Transactions on Graphics (TOG) **32**(6), 195 (2013) [3](#)
5. Chen, W., Gao, J., Ling, H., Smith, E., Lehtinen, J., Jacobson, A., Fidler, S.: Learning to predict 3d objects with an interpolation-based differentiable renderer. In: Advances In Neural Information Processing Systems (2019) [3](#)
6. Choi, S., Zhou, Q.Y., Miller, S., Koltun, V.: A large dataset of object scans. arXiv:1602.02481 (2016) [2](#), [3](#), [12](#)
7. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: European conference on computer vision. pp. 628–644. Springer (2016) [3](#)
8. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR (2017) [2](#), [8](#), [9](#)
9. Delanoy, J., Aubry, M., Isola, P., Efros, A.A., Bousseau, A.: 3d sketching using multi-view deep volumetric prediction. Proceedings of the ACM on Computer Graphics and Interactive Techniques **1**(1), 21 (2018) [3](#), [12](#)
10. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 605–613 (2017) [3](#)
11. Funkhouser, T., Min, P., Kazhdan, M., Chen, J., Halderman, A., Dobkin, D., Jacobs, D.: A search engine for 3d models. ACM Transactions on Graphics (TOG) **22**(1), 83–105 (2003) [3](#)
12. Gao, J., Tang, C., Ganapathi-Subramanian, V., Huang, J., Su, H., Guibas, L.J.: Deepspine: Data-driven reconstruction of parametric curves and surfaces. arXiv preprint arXiv:1901.03781 (2019) [3](#)
13. Gingold, Y., Igarashi, T., Zorin, D.: Structured annotations for 2d-to-3d modeling. In: ACM Transactions on Graphics (TOG). vol. 28, p. 148. ACM (2009) [3](#)
14. Gkioxari, G., Malik, J., Johnson, J.: Mesh r-cnn. arXiv preprint arXiv:1906.02739 (2019) [3](#), [8](#), [14](#)
15. Groueix, T., Fisher, M., Kim, V.G., Russell, B., Aubry, M.: AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2018) [14](#)
16. Guo, R., Hoiem, D.: Support surface prediction in indoor scenes. In: ICCV. pp. 2144–2151 (2013) [2](#)
17. Huang, X., Wang, P., Cheng, X., Zhou, D., Geng, Q., Yang, R.: The apollo-scape open dataset for autonomous driving and its application. arXiv:1803.06184 (2018) [2](#)
18. Isakov, K., Burkov, E., Lempitsky, V., Malkov, Y.: Learnable triangulation of human pose. In: International Conference on Computer Vision (ICCV) (2019) [12](#)
19. J., K., Smith, E., Lafleche, J.F., Fuji Tsang, C., Rozantsev, A., Chen, W., Xiang, T., Lebedev, R., Fidler, S.: Kaolin: A pytorch library for accelerating 3d deep learning research. arXiv:1911.05063 (2019) [3](#)

20. Lee, J., Funkhouser, T.A.: Sketch-based search and composition of 3d models. In: SBM. pp. 97–104 (2008) [3](#)
21. Lim, J.J., Pirsiavash, H., Torralba, A.: Parsing ikea objects: Fine pose estimation. In: ICCV (2013) [2](#)
22. Ling, H., Gao, J., Kar, A., Chen, W., Fidler, S.: Fast interactive object annotation with curve-gcn. In: CVPR. pp. 5257–5266 (2019) [4](#), [8](#), [10](#)
23. Lipson, H., Shpitalni, M.: Conceptual design and analysis by sketching. AI EDAM **14**(5), 391–401 (2000) [3](#)
24. Liu, J., Yu, F., Funkhouser, T.: Interactive 3d modeling with a generative adversarial network. In: 2017 International Conference on 3D Vision (3DV). pp. 126–134. IEEE (2017) [3](#)
25. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4460–4470 (2019) [3](#), [14](#)
26. Oh, S.W., Lee, J., Xu, N., Kim, S.J.: Fast user-guided video object segmentation by interaction-and-propagation networks. CoRR **abs/1904.09791** (2019), <http://arxiv.org/abs/1904.09791> [5](#)
27. Shtof, A., Agathos, A., Gingold, Y., Shamir, A., Cohen-Or, D.: Geosemantic snapping for sketch-based modeling. In: Computer graphics forum. vol. 32, pp. 245–253. Wiley Online Library (2013) [3](#)
28. Smith, E., Fujimoto, S., Meger, D.: Multi-view silhouette and depth decomposition for high resolution 3d object representation. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31, pp. 6479–6489. Curran Associates, Inc. (2018) [3](#)
29. Sun, X., Wu, J., Zhang, X., Zhang, Z., Zhang, C., Xue, T., Tenenbaum, J.B., Freeman, W.T.: Pix3d: Dataset and methods for single-image 3d shape modeling. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) [2](#), [8](#), [9](#), [11](#), [12](#)
30. Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y.G.: Pixel2mesh: Generating 3d mesh models from single rgb images. In: ECCV (2018) [3](#), [8](#), [14](#)
31. Wang, Z., Acuna, D., Ling, H., Kar, A., Fidler, S.: Object instance annotation with deep extreme level set evolution. In: CVPR (2019) [4](#)
32. Xiang, Y., Kim, W., Chen, W., Ji, J., Choy, C., Su, H., Mottaghi, R., Guibas, L., Savarese, S.: Objectnet3d: A large scale database for 3d object recognition. In: ECCV (2016) [2](#)
33. Xie, X., Xu, K., Mitra, N.J., Cohen-Or, D., Gong, W., Su, Q., Chen, B.: Sketch-to-design: Context-based part assembly. In: Computer Graphics Forum. vol. 32, pp. 233–245. Wiley Online Library (2013) [3](#)
34. Xu, Q., Wang, W., Ceylan, D., Mech, R., Neumann, U.: Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. arXiv preprint arXiv:1905.10711 (2019) [3](#)
35. Zhang, X., Zhang, Z., Zhang, C., Tenenbaum, J.B., Freeman, W.T., Wu, J.: Learning to Reconstruct Shapes from Unseen Classes. In: NeurIPS (2018) [3](#), [4](#), [8](#)
36. Zhu, J.Y., Krähenbühl, P., Shechtman, E., Efros, A.A.: Generative visual manipulation on the natural image manifold. In: Proceedings of European Conference on Computer Vision (ECCV) (2016) [3](#)