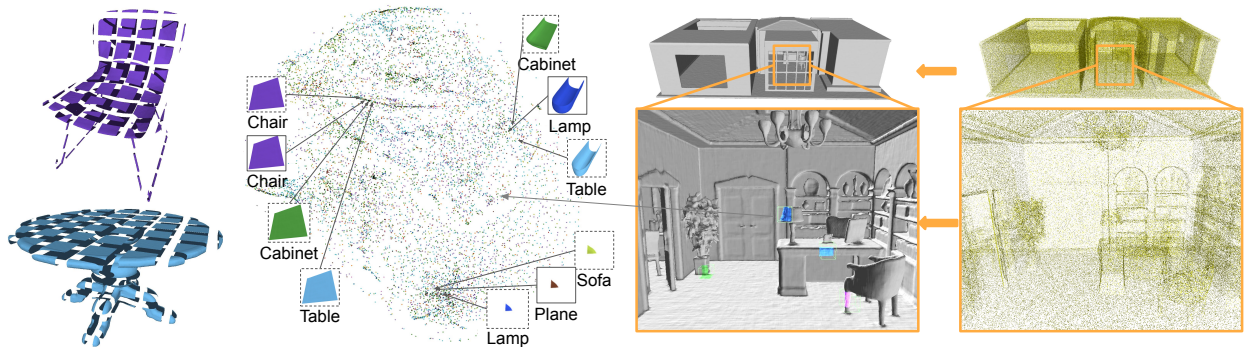
Local Implicit Grid Representations for 3D Scenes

Chiyu "Max" Jiang[1,2] Avneesh Sud[2] Ameesh Makadia[2] Jingwei Huang[2,3]
Matthias Nießner[4] Thomas Funkhouser[2]

[1]UC Berkeley [2]Google Research [3]Stanford University [4]Technical University of Munich



(a) Training parts from ShapeNet. (b) t-SNE plot of part embeddings. (c) Reconstructing entire scenes with Local Implicit Grids

Figure 1: We learn an embedding of parts from objects in ShapeNet [3] using a part autoencoder with an implicit decoder. We show that this representation of parts is generalizable across object categories, and easily scalable to large scenes. By localizing implicit functions in a grid, we are able to reconstruct entire scenes from points via optimization of the latent grid.

Abstract

Shape priors learned from data are commonly used to reconstruct 3D objects from partial or noisy data. Yet no such shape priors are available for indoor scenes, since typical 3D autoencoders cannot handle their scale, complexity, or diversity. In this paper, we introduce Local Implicit Grid Representations, a new 3D shape representation designed for scalability and generality. The motivating idea is that most 3D surfaces share geometric details at some scale – i.e., at a scale smaller than an entire object and larger than a small patch. We train an autoencoder to learn an embedding of local crops of 3D shapes at that size. Then, we use the decoder as a component in a shape optimization that solves for a set of latent codes on a regular grid of overlapping crops such that an interpolation of the decoded local shapes matches a partial or noisy observation. We demonstrate the value of this proposed approach for 3D surface reconstruction from sparse point observations, showing significantly better results than alternative approaches.

1. Introduction

Geometric representation for scenes has been central to various tasks in computer vision and graphics, including geometric reconstruction, compression, and higher-level tasks such as scene understanding, object detection and segmentation. An effective representation should generalize well across a wide range of semantic categories, scale efficiently to large scenes, exhibit a rich expressive capacity for representing sharp features and complex topologies, and at the same time leverage learned geometric priors acquired from data.

In the last years, several works have proposed new network architectures to allow conventional geometric representations such as point clouds [31, 13, 43], meshes [37, 15], and voxel grids [9, 40] to leverage data priors. More recently, a neural implicit representation [4, 28, 29] has been proposed as an alternative to these approaches for its expressive capacity for representing fine geometric details. However, the aforementioned works focus on learning representations for whole objects within one or a few categories, and they have not been studied in the context of generalizing to other categories, or scaling to large scenes.

In this paper we propose a learned 3D shape representation that generalizes and scales to arbitrary scenes. Our key observation is that although different shapes across different categories and scenes have vastly different geometric forms and topologies on a global scale, they share similar features at a certain local scale. For instance, sofa seats and car windshields have a similar curved parts, tabletops and airplane wings both have thin sharp edges, etc.. While no two shapes are the same at the macro scale, and all shapes on a micro-scale can be locally approximated by an angled plane, there exists an intermediate scale (a "part scale"), where a meaningful shared abstraction for all geometries can be learned by a single deep neural network. We aim to learn shape priors at that scale and then leverage them in a scalable and general 3D reconstruction algorithm.

To this end, we propose the Local Implicit Grid (LIG) representation, a regular grid of overlapping part-sized local regions, each encoded with an implicit feature vector. We learn to encode/decode geometric parts of objects at a part scale by training an implicit function autoencoder on 13 object categories from ShapeNet [3]. Then, armed with the pretrained decoder, we propose a mechanism to optimize for the Latent Implicit Grid representation that matches a partial or noisy scene observation. Our representation includes a novel overlapping latent grid mechanism for confidence-weighted interpolation of learned local features for seamlessly representing large scenes. We illustrate the effectiveness of this approach by targeting the challenging application of scene reconstruction from sparse point samples, where we are able to faithfully reconstruct entire scenes given only sparse point samples and shape features learned from ShapeNet objects. Such an approach requires no training on scene level data, where data is costly to acquire. We achieve significant improvement both visually and quantitatively in comparison to state-of-the-art reconstruction algorithms for the scene reconstruction from point samples task (Poisson Surface Reconstruction [23, 24], or PSR, among other methods).

In summary, the main contributions of this work are:

- We propose the Local Implicit Grid representation for geometry, where we learn and leverage geometric features on a part level, and associated methods such as the overlapping latent grid mechanism and latent grid optimization methods for representing and reconstructing scenes at high fidelity.
- We illustrate the significantly improved generalizability of our part-based approach in comparison to related methods that learn priors for entire objects – i.e., we can reconstruct shapes from novel object classes after training only on chairs, or construct entire scenes after training only on ShapeNet parts.
- We apply our novel shape representation approach towards the challenging task of scene reconstruction from sparse point samples, and show significant improvement over the state-of-the-art approach (For Matterport reconstruction from 100/m^2 input points, an F-Score of 0.889 versus 0.455.

2. Related Work

2.1. Geometric representation for objects

In computer vision and graphics, geometric representations such as simplicial complexes (point clouds, line meshes, triangular meshes, tetrahedral meshes) have long been used for representing geometries for its flexibility and compactness. In recent years, various neural architectures have been proposed for analyzing or generating such representations. For instance for [31, 38] have been proposed for analyzing point cloud representations, and [13, 43] for generating point clouds. [27, 17, 20, 19] have been proposed for analyzing signals on meshes, and [37, 15, 7] for generating mesh representations. [21] proposed a general framework for analyzing arbitrary simplicial complex based geometric signals. Naturally paired with 3D Convolutional Neural Networks (CNNs), voxel grids have also been extensively used as a 3D representation [41, 8, 5].

More recently, alternative representations have been proposed in the context of shape generation. Most related to our method are [28, 29, 4], where the implicit surfaces of geometries are represented as spatial functions using fully-connected neural networks. Continuous spatial coordinates are fed as input features to the network which directly produces the values of the implicit functions, however these methods encode the entire shape using a global latent code. [33] used such implicit networks to represent neural features instead of occupancies that can be combined with a differentiable ray marching algorithm to produce neural renderings of objects. Rather than learning a single global implicit network to represent the entire shape, [32] learns a continuous per-pixel occupancy and color representation using implicit networks. Other novel geometric representations in the context of shape reconstruction include Structured Implicit Functions that serves as learned local shape templates [14], and CvxNet [10] which represents space as a convex combination of half-planes that are localized in space. These methods represent entire shapes using a single global latent vector, which can be decoded into continuous outputs with the associated implicit networks.

2.2. Localized geometric representations

Though using a single global latent code to represent entire geometries and scenes is appealing for its simplicity, it fails to capture localized details, and scales poorly to large scenes with increased complexities. [42] proposes to address the localization problem in the context of image to 3D

reconstruction by first estimating a camera pose for the images followed by the projection of local 2D features to be concatenated with global latents for decoding. However, the scalability of such hybrid representations beyond single objects has yet to be shown. Similar to our approach, [39] uses a local patch based representation. However it is not trained on any data, hence is not able to leverage any shape priors from 3d datasets. [30] combines shape patches extracted directly from a set of examples, which limits the shape expressibility. Similar to our spatial partitioning of geometries into part grids, [36] uses PCA-based decomposition to learn a reduced representation of geometric parts within TSDF grids of a fixed scale for the application of real-time geometry compression. These methods do not support scalable reconstruction with learned deep implicit functions.

2.3. Scene-level geometry reconstruction

Most deep learning studies have investigated object reconstruction, with input either as an RGB/D image [5, 37, 28, 4, 13, 10, 14] or 3D points [29, 26, 22], and yet few have considered learning to reconstruct full scenes. Scene level geometry reconstruction is a much more challenging task in comparison to single objects. [34] performs semantic scene completion within the frustum of a single depth image. [8] uses a 3D convolutional network with a coarse-to-fine inference strategy to directly regress gridded Truncated Signed Distance Function (TSDF) outputs from incomplete input TSDF. [1] tackles the scene reconstruction problem by CAD model retrieval, which produces attractive surfaces, at the expense of geometric inaccuracies. However, all of the methods require training on reliable and high-quality scene data. Though several real and synthetic scene datasets exist, such as SunCG [35], SceneNet [16], Matterport3D [2], and ScanNet [6], they are domain-specific and acquiring data for new scenes can be costly. In contrast to methods above that require training on scene dataset, our method naturally generalizes shape priors learned from object datasets and does not require additional training on scenes.

3. Methods

3.1. Method overview

We present a schematic overview of our method in Figure 1. We first learn an embedding of shape parts at a fixed scale from objects in a synthetic dataset using *part autoencoders* (see Sec. 3.2). We show two interesting properties of such a latent embedding: (1) objects that originated from different categories share similar part geometries, validating the generalizability of such learned representations, and (2) parts that are similar in shape are close in the latent space. In order to scale to scenes of arbitrary sizes, we introduce an overlapping gridded representation that can layout these local representations in a scene (Sec. 3.3). Using such part

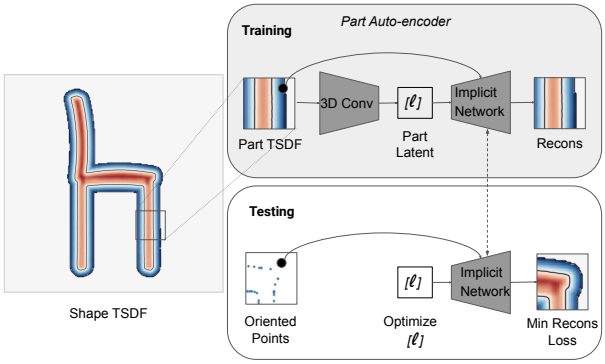


Figure 2: A schematic of the part autoencoder. At train time, crops of the TSDF grid from the ShapeNet dataset are used to train a part autoencoder, with a 3D CNN encoder and implicit network decoder. Interior and exterior points are sampled to supervise the network during training. At inference time, the pre-trained implicit network is attached to a Local Implicit Grid, and the corresponding latent values are optimized via gradient descent on observed interior/exterior points.

embeddings that can be continuously decoded spatially using a local implicit network, we are able to faithfully reconstruct geometries from only sparse oriented point samples by searching for a corresponding latent code using gradient descent-based optimization to match given observations (Sec. 3.4), thus efficiently leveraging geometric priors learned from parts from the ShapeNet dataset.

3.2. Learning a latent embedding for parts

Data Our part embedding model is learned from a collection of 20 million object parts culled from 3D-R^{2}N^2 [5], a 13-class subset of ShapeNet. As preprocessing, we normalize watertight meshes (generated with tools from [28]) into a $[0, 1]$ unit cube, leaving a margin of 0.1 at each side. To maintain the fidelity of the parts, we compute a signed distance function (SDF) at a grid resolution of 256^3. Starting from the origin and with a stride of 16, all 32^3 patches that have at least one point within $3/255$ of the shape surface are extracted as parts for training.

Part Autoencoder We use a 3D CNN decorated with residual blocks for encoding such local TSDF grids, and a reduced IM-NET [4] decoder for reconstructing the part (See Fig. 2). An IM-NET decoder is a simple fully connected neural network with internal skip connections that takes in a latent code concatenated with a 3D point coordinate, and outputs the corresponding implicit function value at the point. We train the network using point samples with binary in/out labels so that the network learns a continuous decision boundary of the binary classifier as the encoded

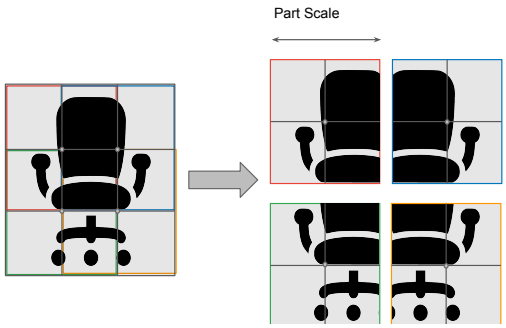


Figure 3: 2D schematic for representing geometries with overlapping latent grids. The implicit value at any point is an bilinear/trilinear interpolation of implicit values acquired by querying 4/8 (2D/3D) neighbors with respect to each cell center.

surface. Since decoding a part is a much more simplified task than decoding an entire shape, we reduce the number of feature channels in each hidden layer of IM-NET by 4 fold, obtaining a leaner and more efficient decoder. To acquire a compact latent representation of parts, we further reduce the number of latent channels for each part to 32. We train the part autoencoder with 2048 random point samples that we sample from the SDF grid on-the-fly during training, where we sample points farther from the boundary with Gaussian-decaying probabilities. The sign of the sample points is interpolated from the sign of the original SDF grid. Furthermore, we truncate the input SDF grids to a value of $3/255$ and renormalize the grid to $[0, 1]$ for stronger gradients near the boundary.

We train the part autoencoder with binary cross entropy loss on the point samples, with an additional latent regularization loss to constrain the latent space of the learned embeddings. The loss is given as:

$$\mathcal{L}(\theta_e, \theta_d) = \frac{1}{|\mathcal{P}||\mathcal{B}|} \sum_{i \in \mathcal{P}} \sum_{j \in \mathcal{B}} \mathcal{L}_c(D_{\theta_d}(\boldsymbol{x}_{i,j}, E_{\theta_e}(g_i)), \text{sign}(\boldsymbol{x}_{i,j})) + \lambda ||E_{\theta_e}(g_i)||_2 \quad (1)$$

where $\mathcal{P}$ is the set of all training parts in a given mini-batch, $\mathcal{B}$ is the set of point samples sampled per part, $\mathcal{L}_c(\cdot, \cdot)$ is the binary cross-entropy loss with logits, E_{θ_e} is the convolutional encoder parameterized by trainable parameters θ_e, D_{θ_d} is the implicit decoder parameterized by trainable parameters θ_d, and g_i is the input tsdf grid for the i-th part, $\text{sign}(\cdot)$ takes the sign of the corresponding point $x_{i,j}$.

3.3. Local implicit grids

In order to use the learned part representations for representing entire objects and scenes, we lay out a sparse latent grid structure, where within each local grid cell the surface is continuously decoded from the local latent codes within the cell. In world coordinates, when querying for the implicit function value at location $\boldsymbol{x}$ against a single voxel grid cell centered at $\boldsymbol{x_i}$, the implicit value is decoded as:

$$f(\boldsymbol{x}, \boldsymbol{c}_i) = D_{\theta_d}(\boldsymbol{c}_i, \frac{2}{s}(\boldsymbol{x} - \boldsymbol{x}_i)) \quad (2)$$

where $\boldsymbol{c}_i$ is the latent code corresponding to the part in cell i, and s is the part scale. The coordinates are first being transformed into normalized local coordinates within the cell to $[-1, 1]$, before being queried against the decoder.

Though directly partitioning space into a voxel grid with latent channels within each cell gives decent performance, there will be discontinuities across voxel boundaries. Hence we propose the overlapping latent grid scheme, where each grid cell for a part overlaps with its neighboring cells by half the part scale (see Fig. 3). When querying for the implicit function value at an arbitrary position $\boldsymbol{x}$ against overlapping latent grids, the value is computed as a trilinear interpolation of independent queries to all cells that overlap at this position, which is 4 in 2 dimensions and 8 in 3 dimensions:

$$f(\boldsymbol{x}, \{\boldsymbol{c}_j | j \in \mathcal{N}\}) = \sum_{j \in \mathcal{N}} w_j D_{\theta_d}(\boldsymbol{c}_j, \frac{2}{s}(\boldsymbol{x} - \boldsymbol{x}_j)) \quad (3)$$

where $\mathcal{N}_j$ is the set of all neighboring cells of point $\boldsymbol{x}$, and w_j is the trilinear interpolation weight corresponding to cell j. Under such an interpolation scheme, the overall function represented by the implicit grid is guaranteed to be C^0 continuous. Higher-order continuity could be similarly acquired with higher degrees of polynomial interpolations, though we do not explore it in the scope of this study. For additional efficiency, since most grid cells do not have any points that fall into them, we use a sparse data structure for storing latent grid values, optimization, and decoding for the reconstructed surface, where empty space is assumed to be exterior space.

3.4. Geometric encoding via latent optimization

At inference time, when presented with a sparse point cloud of interior/exterior samples as input, we decompose space into a coarse grid and then perform optimization for the latent vectors associated with the grid cells in order to minimize the cost function for classifying sampled interior/exterior points. The initial values within the latent grid is initialized as random normal with a standard deviation of 10^{-2}. If we denote the set of effective latent grid cells as $\mathcal{G}$, the corresponding latent code in each grid cell c_j, and the set of all sampled interior/exterior input points as $\mathcal{B}$, we optimize the latent codes for the minimal classification loss on the sampled points:

$$\arg\min_{\boldsymbol{c} \in \mathcal{G}} \sum_{i \in \mathcal{B}} \sum_{j \in \mathcal{N}_i} \mathcal{L}_c(f(\boldsymbol{x_i}, \{\boldsymbol{c}_j | j \in \mathcal{N}\}), \text{sign}(\boldsymbol{x}_i)) + \lambda ||\boldsymbol{c}_j||_2 \quad (4)$$

How do we acquire the signed point samples for performing this latent grid optimization? For autoencoding a

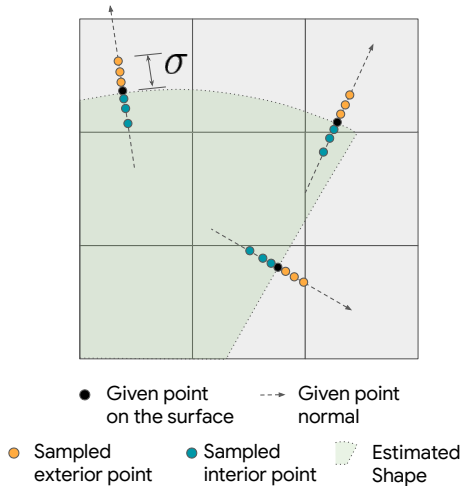


Figure 4: Schematic for reconstructing shapes based on sparse oriented point samples. Given original point samples on the surface with normals, we randomly sample k samples along both sides of each normal vector and assign signs for these samples accordingly. The points are sampled with a Gaussian falloff probability, with a given standard deviation σ. The latent codes within the overlapping latent grids are updated via optimization for minimizing classification loss as in Eqn. 4. The surface of the shape is reconstructed by densely querying the latent grid and extracting the zero-contour of the output logits.

geometry with a latent grid, the signed point samples are densely sampled near the surface of the given shape to be encoded. However, for the application of recovering surface geometry from sparse oriented point samples, we randomly sample interior and exterior points for each point sample along the given normal direction, with a Gaussian falloff probability parameterized by a standard deviation of σ. See Fig. 4 for details. All grid cells that do not contain any point from the input point cloud is assumed to be an empty exterior volume. This is effective and works well for scenes that do not contain large enclosed volumes, but creates an artificial back-face in the enclosed interior. We detail a simple postprocessing algorithm to remove such artifacts resulting from the exterior empty space assumption. We provide more details about the additional postprocessing algorithm in the Appendix.

As our method requires optimizing over the learned latent space, it is reasonable to wonder if alternate models such as a variational autoencoder [25] or autodecoder [29] would be a more appropriate choice, as both formulations incorporate a latent distribution prior. However, [29] observed the stochastic nature of the VAE made training difficult. Also, the autodecoder is fundamentally unable to scale to large numbers of parts at training as it requires fast storage and random access to all latent embeddings during training. These concerns motivated our decision to adopt an autoencoder formulation with a regularization loss to constrain the latent space.

4. Experiments

We ran a series of experiments to test the proposed LIG method. We focus on two properties of our method: the generalization of our learned part representation, and the scalability of our learned shape representation to large scenes. Our target application is reconstructing scenes from a sparse set of oriented point samples, a challenging task that requires learned part priors for detailed and accurate reconstruction.

Metrics In all of our experiments, we evaluate geometric reconstruction quality with Chamfer Distance (CD), Normal Alignment (Normal), and F-Score. For Chamfer Distance and Normal Alignment, we base our implementation on [28] with small differences. For object-level autoencoding experiments, we follow [13, 28] and normalize the unit distance to be 1/10 of the maximal edge length of the current objects bounding box. We estimate CD and Normal Alignment using 100,000 randomly sampled points on the ground truth and reconstructed meshes. For the two scene-level experiments, we randomly sample 2 million points on each mesh when estimating CD and Normal Alignment. When evaluating scene reconstructions, we use world coordinate scales (meters) for computing CD, since data is provided in a physically-meaningful scale. Additionally, in all experiments, we compute the F-Score at a threshold of τ, as F-Score is a metric less sensitive to outliers. F-Score is the harmonic mean of recall (percentage of reconstruction to target distances under τ) and precision (vice versa). For object reconstruction (Sec. 3.2) we use $\tau = 0.1$ and for scene reconstruction, we use $\tau = 0.025$ (i.e., 2.5cm).

4.1. Generalization of learned part representation

Task In order to investigate the generalization of the learned embedding by reducing the scale of the learned shape from object scale to part scale, we construct an investigative experiment of training the models to learn a shape autoencoder on a single category of objects (in this case, chairs in the training set of ShapeNet), and reconstructing examples from the all 13 object categories, including the other 12 unseen categories.

Baseline As our main objective is to explore the gain in generalizability from learning an embedding of part scales, we benchmark our method against the original IM-NET decoder with a similar 3D convolution based encoder as the

Category	IM-NET			Ours		
	CD (↓)	Normal (↑)	F-Score (↑)	CD (↓)	Normal (↑)	F-Score (↑)
chair	0.181	0.820	0.505	**0.099**	**0.920**	**0.710**
airplane	0.698	0.550	0.151	0.150	0.817	0.564
bench	0.229	0.719	0.433	0.054	0.905	0.857
cabinet	0.343	0.700	0.230	0.118	0.948	0.733
car	0.354	0.646	0.240	0.152	0.825	0.472
display	0.601	0.574	0.130	0.170	0.926	0.551
lamp	0.836	0.592	0.120	0.114	0.882	0.624
loudspeaker	0.377	0.702	0.246	0.139	0.937	0.711
rifle	0.902	0.400	0.080	0.113	0.824	0.693
sofa	0.199	0.812	0.484	0.077	0.944	0.822
table	0.425	0.681	0.242	0.066	0.936	0.844
telephone	0.623	0.547	0.120	0.037	0.984	0.962
vessel	0.591	0.574	0.147	0.178	0.847	0.467
mean*	0.435	0.666	0.274	**0.114**	**0.898**	**0.692**

Table 1: Shape autoencoding for autoencoders trained on only chairs and evaluated on all 13 categories. The mean corresponds to class-averaged mean of all out-of-training object categories.

Metrics	CD(↓)	Normal(↑)	F-Score(↑)
IM-NET	0.183	0.827	0.647
Ours	**0.007**	**0.945**	**0.985**

Table 2: Qualitative comparison of scene representational performance for IM-NET versus our method.

encoder part of our part autoencoder. To implement autoencoding for our method, we train our autoencoder on all the parts we extract from the training split of the chair category in ShapeNet. We then "encode" the geometries of the unseen shapes using the latent optimization method that is described in Sec. 3.4.

Results Discussion We quantitatively and qualitatively compare reconstruction performances in Table 1 and Figure 5, respectively. Given an IM-NET that is trained to learn a latent representation of objects (in this scenario, chairs), the learned representation does not generalize to classes beyond the source class. Visually, IM-NET achieves good reconstructions on the source class as well as related classes (e.g., sofa), but performs poorly on semantically different classes (e.g., airplane). In contrast, the part representation learned by our local implicit networks is transferable across drastically different object categories.

4.2. Scalability of scene representational power

Task As a second experiment, we investigate the increased representational power and scalability that we gain from learning a part-based shape embedding. The definition of the task is: given one scene, what is the best reconstruction performance we can get from either representation for memorizing and overfitting to the scene.

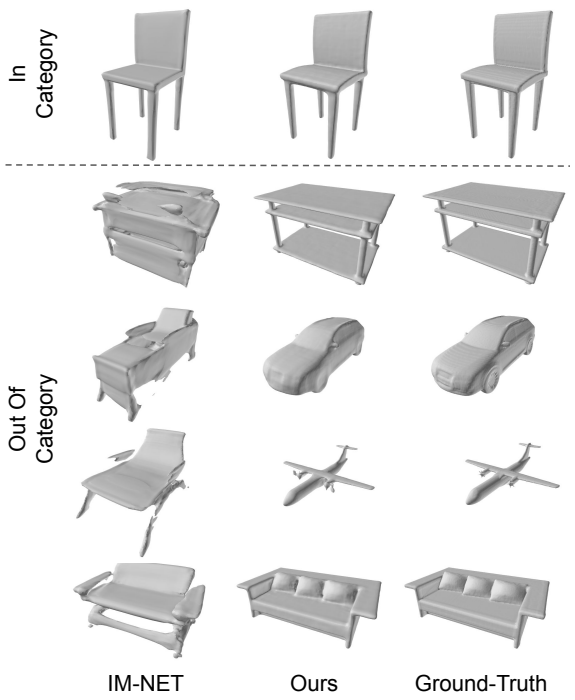


Figure 5: Qualitative comparison of autoencoded shape from in-category (chair) and out-of-category shapes. IM-NET trained to learn embeddings of one object category does not transfer well to unseen categories, while the part embedding learned by our local implicit networks is much more transferable across unseen categories.

Baseline Similar to the previous experiment, we compare directly with IM-NET for representational capacity towards a scene, as it is the decoder backbone that our method is based on, to investigate the improvement in scalability that we are able to gain by distributing geometric information in spatially localized grid cells versus a single global representation. For this task, as the objective is to encode one scene, we use the encoderless version of IM-NET, where during training time, the decoder only receives spatial coordinates of point samples (not concatenated with a latent code) that are paired with the signs of these points. For our method, we use latent optimization against the pretrained decoder for encoding the scenes, using 100k surface point samples from the scene, with a sampling factor of $k = 10$ per point along the normal direction.

Data We evaluate the representational qualities of the two methods on the meshes from the validation set of the Matterport 3D [2] scene dataset. We perform the evaluations at the region level of the dataset, requiring the models to encode one region at a time. Additionally, we provide one example from SceneNet for visual comparison in Fig. 6.

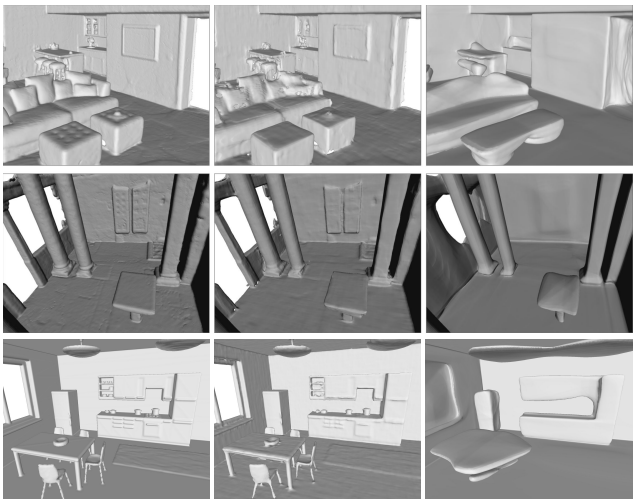


Figure 6: Qualitative comparison of the scene representational performance: Left to right: Ground truth scene, our reconstruction using sampling density 500 points/m^2, and IM-NET. First two rows from Matterport, last row from SceneNet.

Results Discussion The quantitative (Table 2) and qualitative (Fig. 6) results are presented. While IM-NET is able to reconstruct the general structure of indoor scenes such as smooth walls and floors, it fails to capture fine details of objects due to the difficulty of scaling a single implicit network to an entire scene. Our Local Implicit Grids are able to capture global structures as well as local details.

4.3. Scene reconstruction from sparse points

Task As a final task and our main application, we apply our reconstruction method to the classic task in computer graphics to reconstruct geometries from sparse points. This is an important application since surface reconstruction from points is a crucial step in the process of digitizing the 3-dimensional world. The input to the reconstruction pipeline is the sparse point samples that we randomly sample from the surface mesh of the scene datasets. We study reconstruction performances with a varied number of input point samples and point densities.

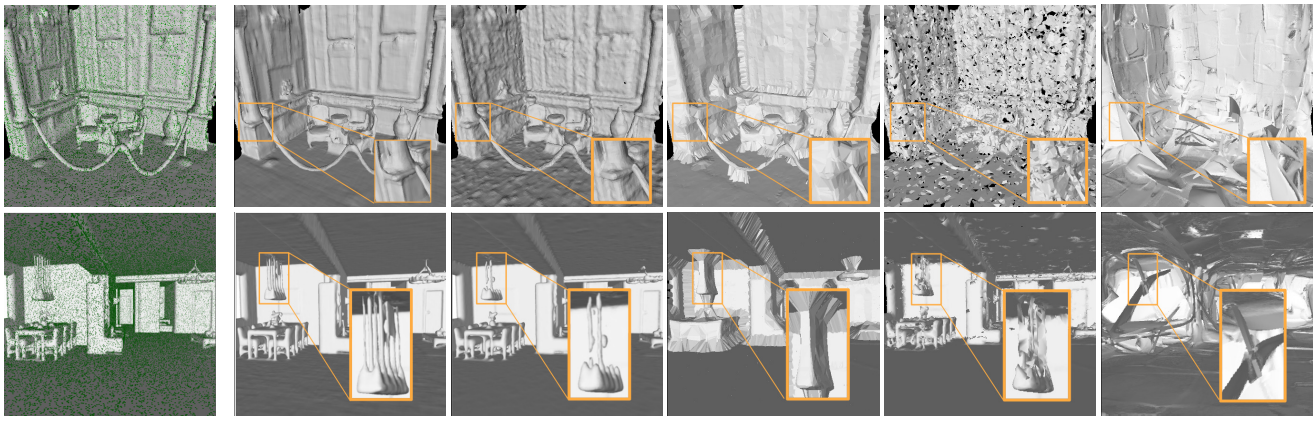
Baseline We mainly compare our method to the traditional Poisson Surface Reconstruction (PSR) method [23, 24] with a high octree depth value (depth=10) for the scene reconstruction experiment, which remains the state-of-the-art method for surface reconstruction tasks of scenes. We also compare with other classic (PSR at depth 8 and 9, Alpha Complex [11], Ball Pivoting [12]) and deep (Deep Geometric Prior [39]) reconstruction methods on one representative scenario (see 100pts/m^2 in Table 3) due to the high computational cost of evaluating all methods on all scenes. While various other deep learning based methods [29, 26, 22] have been proposed for surface reconstruction from points in a similar setting, all of the deep learning based methods are object-specific, trained and tested on specific object categories in ShapeNet, with no anticipated transferability to unseen categories or scenes, as we have shown in the experiment in Sec. 4.1. Furthermore, as both PSR and our method require no training/finetuning on the scene level datasets, the task is based on the premise that high-quality 3D training data is costly to acquire or unavailable for scenes. For our method, we adaptively use different part sizes for different point densities. We use 25cm (1000 pts/m^2), 35cm (500 pts/m^2), 50cm (100 pts/m^2) and 75cm (20 pts/m^2) corresponding to different point densities for optimal performance.

Data We evaluate the reconstruction performance of the methods on a synthetic dataset: SceneNet [16], and a high quality scanned dataset: Matterport 3D [2] (validation split). As both SceneNet and Matterport 3D datasets are not watertight, and in addition to that, SceneNet dataset has various artifacts such as double-sided faces that produce conflicting normal samples, we preprocess both datasets using the watertight manifold algorithm as describe in [18]. For both datasets, as the scenes vary in sizes, we sample a constant density of points on mesh surfaces (20, 100, 500 and 1000 points per m^2). As preprocessing produces large empty volumes for SceneNet, we drop scenes that have a volume-to-surface-area ratio lower than 0.13.

Results Discussion We compare the reconstruction performances in Table 3 and 4, and Fig. 7. With a high number of input point samples, both PSR10 and our method are able to reconstruct the original scene with high fidelity. However, with a low number of point samples, our method is able to leverage geometric priors to perform a much better reconstruction than PSR. Additionally, our method is able to reconstruct thin structures very well whereas PSR fails to do so. However, since our method only reconstructs finite thickness surfaces as determined by finite part size, it creates double sided surfaces on the enclosed non-visible interiors, leading to degraded performance in F-Score for the 500 and 1000 pts/m^2 scenarios in Table 3.

5. Ablation Study

Additionally, we study the effects of two important aspects of our method: the part scale that we choose for reconstructing each scene, and overlapping latent grids. We choose SceneNet reconstruction from 100 point samples / m^2 as a representative case for the ablation study. See Table 5 for a comparison. As seen from the results, the reconstruction results are affected by the choice of part scale,



GT+Input PC (a) Ours (b) PSR-10 [24] (c) Alpha Cmplx [11] (d) Ball Pivot [12] (e) DGP [39]

Figure 7: Qualitative comparisons of scene reconstruction performance from sparse oriented point samples. Our method is significantly better at reconstructing scenes from sparse point clouds compared to baseline methods, especially with respect to sharp edges and thin structures.

points/m^2	Method	CD(↓)	Normal(↑)	F-Score(↑)
20	PSR10	0.077	0.802	0.317
	Ours	**0.017**	**0.920**	**0.859**
100	PSR8	0.031	0.891	0.721
	PSR9	0.035	0.890	0.721
	PSR10	0.035	0.890	0.725
	Alpha	0.021	0.709	0.736
	BallPvt	0.015	0.880	0.839
	DGP	0.037	0.852	0.571
	Ours	**0.012**	**0.961**	**0.957**
500	PSR10	0.024	0.959	0.957
	Ours	**0.010**	**0.976**	**0.972**
1000	PSR10	0.026	0.975	0.984
	Ours	**0.009**	**0.984**	**0.986**

Table 3: Reconstruction performance on SceneNet dataset.

points/m^2	Method	CD(↓)	Normal(↑)	F-Score(↑)
20	PSR10	0.167	0.655	0.276
	Ours	**0.028**	**0.813**	**0.691**
100	PSR10	0.106	0.757	0.455
	Ours	**0.013**	**0.883**	**0.889**
500	PSR10	0.103	0.871	0.778
	Ours	**0.008**	**0.928**	**0.970**
1000	PSR10	0.102	0.910	0.862
	Ours	**0.007**	**0.945**	**0.985**

Table 4: Reconstruction performance on Matterport dataset.

CL	PS	Overlap	CD(↓)	Normal(↑)	F-Score(↑)
32	25cm	Yes	0.013	0.948	0.921
32	50cm	Yes	0.012	0.961	0.957
32	75cm	Yes	0.013	0.945	0.929
32	50cm	No	0.023	0.886	0.857
8	50cm	Yes	0.016	0.925	0.879

Table 5: Ablation study on the effects of the choice of latent code length (CL), part scale (PS), and overlapping latent grid design on the reconstruction performance for scenes.

albeit not very heavily influenced. Overlapping latent grids significantly improves the quality of the overall reconstruction. With a smaller latent code size of 8, the performance is slightly deteriorated due to more limited expressivity for part geometries.

6. Discussion and Future Work

The Local Implicit Grid (LIG) representation for 3D scenes is a regular grid of overlapping part-sized local regions, each encoded with an implicit feature vector. Experiments show that LIG is capable of reconstructing 3D surfaces of objects from classes unseen in training. Furthermore, to our knowledge, it is the first learned 3D representation for reconstructing scenes from sparse point sets in a scalable manner. Topics for future work include ways to constrain the LIG optimization to produce latent codes near training examples, explore alternate implicit function representations (e.g., OccNet), and to investigate the best ways to use LIG for 3D reconstruction from image(s).

Acknowledgements

We would like to thank Kyle Genova, Fangyin Wei, Abhijit Kundu, Alireza Fathi, Caroline Pantofaru, David Ross, Yue Wang, Mahyar Najibi and Chris Bregler for helpful discussions, Angela Dai for help with supplemental video, JP Lewis for offering help in paper review, as well as anonymous reviewers for helpful feedback. This work was supported by the ERC Starting Grant *Scan2CAD* (804724).

References

[1] Armen Avetisyan, Manuel Dahnert, Angela Dai, Manolis Savva, Angel X Chang, and Matthias Nießner. Scan2cad: Learning cad model alignment in rgb-d scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2614–2623, 2019. 3

[2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 3, 6, 7

[3] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1, 2

[4] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 1, 2, 3, 11

[5] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, pages 628–644. Springer, 2016. 2, 3

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 3

[7] Angela Dai and Matthias Nießner. Scan2mesh: From unstructured range scans to 3d meshes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5574–5583, 2019. 2

[8] Angela Dai, Daniel Ritchie, Martin Bokeloh, Scott Reed, Jürgen Sturm, and Matthias Nießner. Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2018. 2, 3

[9] Angela Dai, Charles Ruizhongtai Qi, and Matthias Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5868–5877, 2017. 1

[10] Boyang Deng, Kyle Genova, Soroosh Yazdani, Sofien Bouaziz, Geoffrey Hinton, and Andrea Tagliasacchi. Cvxnets: Learnable convex decomposition. *arXiv preprint arXiv:1909.05736*, 2019. 2, 3

[11] Edelsbrunner and Mücke. Three-dimensional alpha shapes. *ACM TOG*, 13(1):43–72, 1994. 7, 8

[12] Bernardini et al. The ball-pivoting algorithm for surface reconstruction. *IEEE VCG*, 5(4):349–359, 1999. 7, 8

[13] Haoqiang Fan, Hao Su, and Leonidas J Guibas. 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*, pages 605–613, 2017. 1, 2, 3, 5

[14] Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T Freeman, and Thomas Funkhouser. Learning shape templates with structured implicit functions. *arXiv preprint arXiv:1904.06447*, 2019. 2, 3

[15] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. Atlasnet: A papier-mache approach to learning 3d surface generation. *arXiv preprint arXiv:1802.05384*, 2018. 1, 2

[16] Ankur Handa, Viorica Patraucean, Vijay Badrinarayanan, Simon Stent, and Roberto Cipolla. Understanding real world indoor scenes with synthetic data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4077–4085, 2016. 3, 7

[17] Rana Hanocka, Amir Hertz, Noa Fish, Raja Giryes, Shachar Fleishman, and Daniel Cohen-Or. Meshcnn: a network with an edge. *ACM Transactions on Graphics (TOG)*, 38(4):90, 2019. 2

[18] Jingwei Huang, Hao Su, and Leonidas Guibas. Robust watertight manifold surface generation method for shapenet models. *arXiv preprint arXiv:1802.01698*, 2018. 7

[19] Jingwei Huang, Haotian Zhang, Li Yi, Thomas Funkhouser, Matthias Nießner, and Leonidas J Guibas. Texturenet: Consistent local parametrizations for learning from high-resolution signals on meshes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4440–4449, 2019. 2

[20] Chiyu Jiang, Jingwei Huang, Karthik Kashinath, Philip Prabhat, Marcus, and Matthias Niessner. Spherical cnns on unstructured grids. In *International Conference on Learning Representations*, 2019. 2

[21] Chiyu Jiang, Dana Lynn Ona Lansigan, Philip Marcus, Matthias Nießner, et al. Ddsl: Deep differentiable simplex layer for learning geometric signals. *arXiv preprint arXiv:1901.11082*, 2019. 2

[22] Chiyu Jiang, Dequan Wang, Jingwei Huang, Philip Marcus, Matthias Nießner, et al. Convolutional neural networks on non-uniform geometrical signals using euclidean spectral transformation. *arXiv preprint arXiv:1901.02070*, 2019. 3, 7

[23] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7, 2006. 2, 7

[24] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):29, 2013. 2, 7, 8, 13, 14

[25] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proc. ICLR*, 2014. 5

[26] Yiyi Liao, Simon Donne, and Andreas Geiger. Deep marching cubes: Learning explicit surface representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2916–2925, 2018. 3, 7

[27] Jonathan Masci, Davide Boscaini, Michael Bronstein, and Pierre Vandergheynst. Geodesic convolutional neural networks on riemannian manifolds. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 37–45, 2015. 2

[28] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 1, 2, 3, 5

[29] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 1, 2, 3, 5, 7

[30] Mark Pauly, Niloy J Mitra, Joachim Giesen, Markus H Gross, and Leonidas J Guibas. Example-based 3d scan completion. In *Symposium on Geometry Processing*, number CONF, pages 23–32, 2005. 3

[31] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017. 1, 2

[32] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *arXiv preprint arXiv:1905.05172*, 2019. 2

[33] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *arXiv preprint arXiv:1906.01618*, 2019. 2

[34] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1746–1754, 2017. 3

[35] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. *Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[36] Danhang Tang, Mingsong Dou, Peter Lincoln, Philip Davidson, Kaiwen Guo, Jonathan Taylor, Sean Fanello, Cem Keskin, Adarsh Kowdle, Sofien Bouaziz, et al. Real-time compression and streaming of 4d performances. In *SIGGRAPH Asia 2018 Technical Papers*, page 256. ACM, 2018. 3

[37] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–67, 2018. 1, 2, 3

[38] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 38(5):146, 2019. 2

[39] Francis Williams, Teseo Schneider, Claudio Silva, Denis Zorin, Joan Bruna, and Daniele Panozzo. Deep geometric prior for surface reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10130–10139, 2019. 3, 7, 8

[40] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in neural information processing systems*, pages 82–90, 2016. 1

[41] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 2

[42] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *arXiv preprint arXiv:1905.10711*, 2019. 2

[43] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4541–4550, 2019. 1, 2

Appendix

A. Additional implementation details

A.1. Model architecture

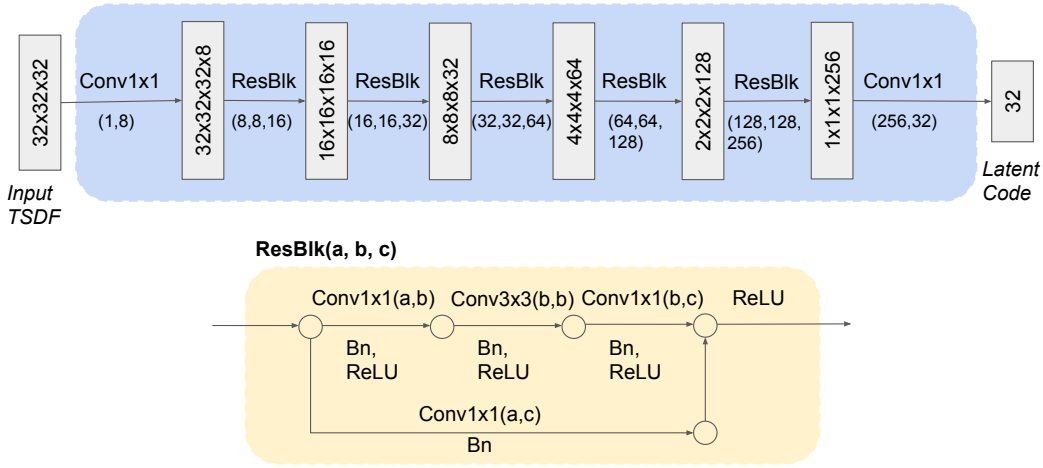


Figure 8: Encoder architecture. The encoder is a simple 3D CNN decorated with residue blocks, that encodes 3D TSDF tensors into latent codes, which can be decoded into implicit surfaces by an implicit network decoder.

We present a schematic of our encoder architecture for our part autoencoder in Fig. 8. The input to the encoder is a normalized TSDF crop of the part to be encoded, and the encoder uses 3D CNNs to encode the input into a latent code of dimensions 32. The encoder is decorated with residue blocks with bottleneck layers for improved performance.

We refer the reader to [4] for the architecture for our refiner. We preserve the architecture of the IM-NET model, but reduce the latent dimension from 128 to 32, and reduce the number of hidden layers in every layer of the model to 1/4 of the original value for improved efficiency, due to the fact that part geometries are easier to learn and represent than entire objects.

A.2. Part autoencoder training

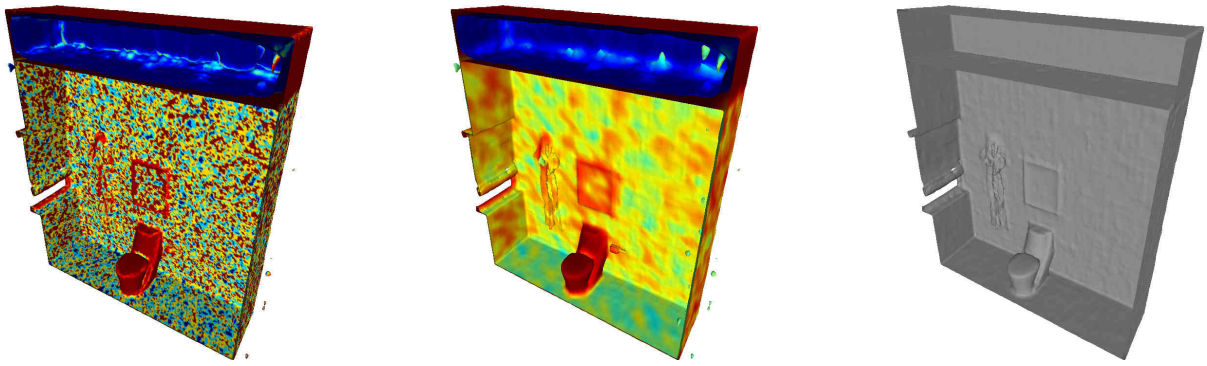
For training the part autoencoder, we use a batch size of 32, and for each shape we sample 2048 point samples. We train with a latent penalty factor $\lambda = 10^{-2}$, learning rate of 10^{-3}. We sample empty volumes with a probability of 10^{-3} to embed empty space. We train the part autoencoder for a total of 10^7 steps.

A.3. Inference

For reconstructing geometries from point samples, for each point sample, we sample 10 points along the point normal with a standard deviation of 1cm. For the Local Implicit Grid, we initialize each cell with Gaussian normal random values with a standard deviation of 0.01. During latent grid optimization, we use 32768 random point samples per batch, and optimize with a learning rate of 10^{-3}. We optimize for a fixed 10000 steps. When extracting the final mesh, we extract the mesh at $1/64m$ resolution.

A.4. Postprocessing algorithm

As discussed in the main text, one undesired side product from assuming all empty LIG grid cells to be "exterior" space is that it results in back-faces enclosed in large volumes. A simple postprocessing algorithm can be devised to remove such artifacts. For every face in the reconstructed mesh, we first compute the centroid of each face, as well as its normal direction. For the centroid of each face, we find the top-k nearest points in the original input oriented point set and compute the dot product of the normals between the pair of points. As such, back-faces will consistently have the opposite sign, and the exterior face will have the correct sign. This, however, will be noisy and non-robust to thin surfaces (with both sides very close to each other), since approximately half of the time the faces will find an input point on the opposite side as its nearest neighbor (see Fig. 9a). This can be effectively mitigated by using a Laplacian kernel (diffusion coefficient λ, i iterations)



(a) Before postprocessing. Color by original mesh normal alignment signal.

(b) Before postprocessing. Color by normal alignment signal after Lap. smoothing.

(c) Postprocessed Reconstructed Mesh

Figure 9: Schematics for postprocessing algorithm. The back-face artifact in the original reconstructed mesh can be clearly seen in dark blue, and is effectively removed in the postprocessed mesh (c).

to smooth the normal alignment signal, followed by discarding all faces below a certain normal alignment threshold n, and discarding all disconnected components with an area below a.

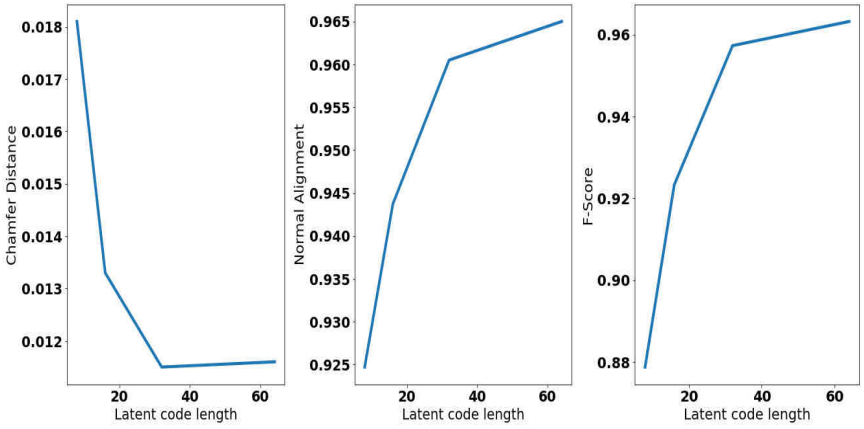
In all our cases, we used the parameters $k = 3, n = -0.75, \lambda = 0.5, i = 50, a = 1$.

B. Additional ablation studies

We perform additional ablation studies on the effects of latent code length on reconstruction performace. See Table 6 and Fig. 10 for reference. With increasing number of latent channels, the reconstruction performance improves with diminishing marginal improvement. Our choice of 32 latent channels strikes a good balance between performance and efficiency.

CL	CD(↓)	Normal(↑)	F-Score(↑)
8	0.018	0.925	0.879
16	0.013	0.944	0.923
32	0.012	0.961	0.957
64	0.012	0.965	0.963

Table 6: Additional ablation study on the effects of latent code length (CL). Reconstruction performance measured on SceneNet reconstruction from 100 point samples / m^2.



(a) Chamfer Distance

(b) Normal Alignment

(c) F-Score

Figure 10: Line plot for Chamfer Distance, Normal Alignment and F-Score versus Latent Code Length.

C. Additional visual results

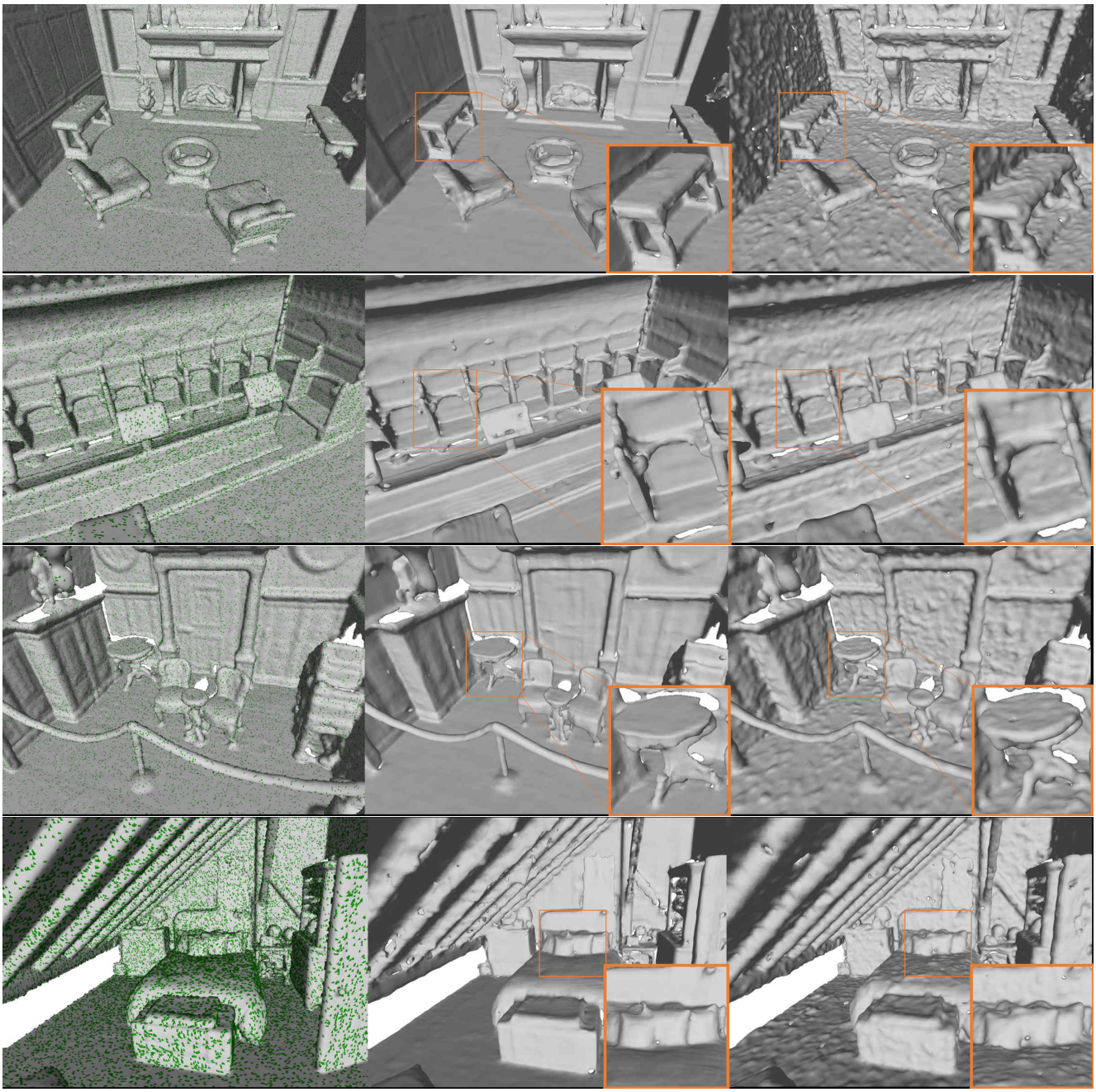


Figure 11: Left: Ground truth mesh overlaid with input point samples; Middle: Our reconstruction; Right: Screened PSR [24] reconstruction. The input are point samples from the Matterport ground truth mesh at a sample density of 500 points / m^2.

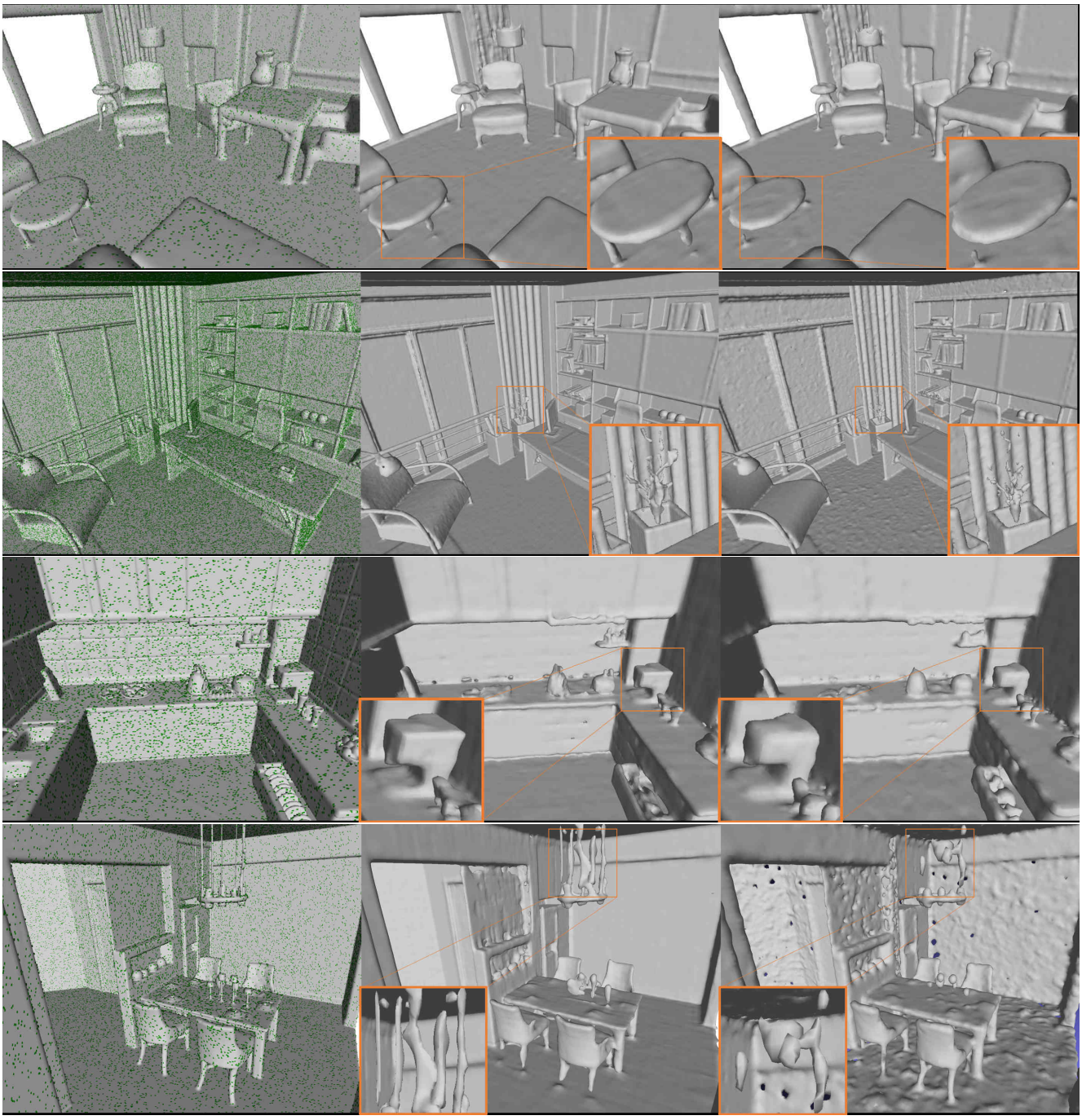


Figure 12: Left: Ground truth mesh overlaid with input point samples; Middle: Our reconstruction; Right: Screened PSR [24] reconstruction. The input are point samples from the SceneNet ground truth mesh at a sample density of 500 points / m^2.