Neural Part Priors: Learning to Optimize Part-Based Object Completion in RGB-D Scans

Aleksei Bokhovkin
Technical University of Munich
aleksei.bokhovkin@tum.de

Angela Dai
Technical University of Munich
angela.dai@tum.de

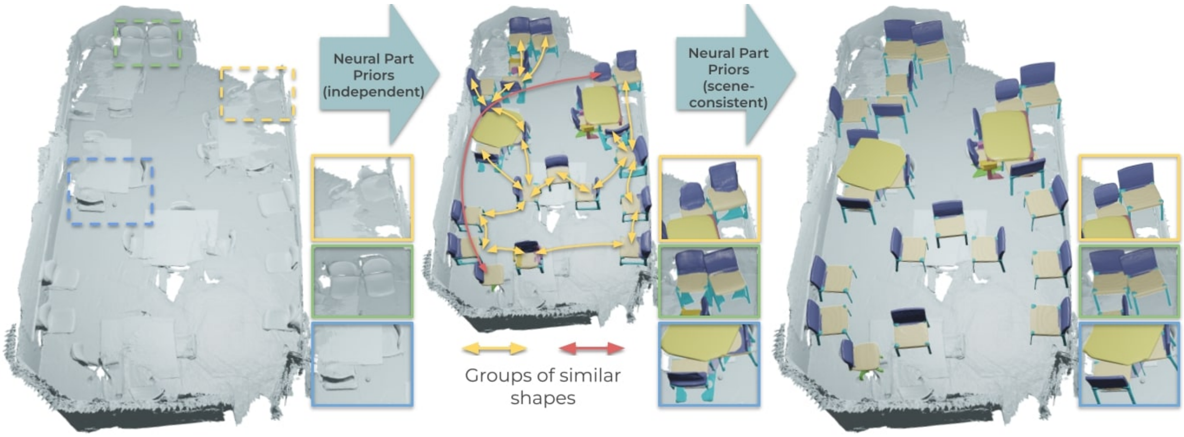

Neural Part Priors (independent)
Neural Part Priors (scene-consistent)
Groups of similar shapes

Figure 1. Our Neural Part Priors learn optimizable parametric latent spaces of object part geometries, which we can use to fit partial, real-world RGB-D scans of a scene, decomposing detected objects into their complete part geometries. In contrast to 3D scene understanding approaches that make independent predictions per-object, our parametric part spaces enables formulating test-time constraints for consistency within an input scene, thus producing both accurate as well as globally-consistent part decompositions.

Abstract

3D scene understanding has seen significant advances in recent years, but has largely focused on object understanding in 3D scenes with independent per-object predictions. We thus propose to learn Neural Part Priors (NPPs), parametric spaces of objects and their parts, that enable optimizing to fit to a new input 3D scan with global scene consistency constraints. The rich structure of our NPPs enables accurate, holistic scene reconstruction across similar objects in the scene. Both objects and their part geometries are characterized by coordinate field MLPs, facilitating optimization at test time to fit to input geometric observations as well as similar objects in the input scan. This enables more accurate reconstructions than independent per-object predictions as a single forward pass, while establishing global consistency within a scene. Experiments on the ScanNet dataset demonstrate that NPPs significantly outperforms the state-of-the-art in part decomposition and object completion in real-world scenes.

1. Introduction

With the introduction of commodity RGB-D sensors (e.g., Microsoft Kinect, Intel RealSense, etc.), remarkable progress has been made in reconstruction and tracking to construct 3D models of real-world environments [9, 14, 37, 39, 52]. This has enabled construction of large-scale datasets of real-world 3D scanned environments [4, 11], enabling significant advances in 3D semantic segmentation [10, 13, 23], 3D semantic instance segmentation [18, 24, 26], and even part-level understanding of scenes [3]. The achieved 3D object recognition has shown impressive advances, but methods focus on independent predictions per object in single forward passes, resulting in semantic predictions that are inconsistent between repeated objects in a scene, and/or geometric predictions that do not precisely match input observed geometry.

Simultaneously, recent advances in representing 3D shapes as continuous implicit functions represented with coordinate field MLPs have shown high-fidelity shape reconstruction [6–8, 40, 47]. Such methods have focused on object-level reconstructions, whereas part-based understanding is fundamental to many higher-level scene understanding tasks (e.g., interactions often occur with object parts – sitting on the seat of a couch, opening a door with a handle, etc.).

We thus propose to learn Neural Part Priors (NPPs), optimizable parametric shape and part spaces learned from synthetic data. These learned manifolds enable efficient traversal in latent spaces during inference to fit precisely to objects in real-world scanned scenes, while maintaining consistent part decompositions with similar objects in the scene. Our NPPs leverage the representation power of neural implicit functions encoded as coordinate-field MLPs, representing both shape and part geometries of objects. A shape can then be represented by a set of latent codes for each of its parts, where each code decodes to predict the respective part segmentation and signed distance field representation of the part geometry. This representation enables effective *test-time joint optimization* over all parts of a shape by traversing through the part latent space to find the set of parts that best explain a real-world shape observation. Furthermore, as repeated objects often appear in a scene under different partial observation patterns (resulting in inconsistent predictions when made independently for each object) we further optimize for part consistency between similar objects detected in a scene to produce scene-consistent part decompositions. This allows us to reconstruct the holistic structure of a scene.

To fit real-world 3D scan data, we first perform object detection and estimate the part types for each detected object. We can then optimize *in test time* jointly over the part codes for each shape to fit the observed scan geometry; we leverage a predicted part segmentation of the detected object and optimize jointly across the parts of each shape such that each part matches the segmentation, and their union fits the object. This joint optimization across parts produces a high-resolution part decomposition whose union represents the complete shape while fitting precisely to the observed real geometry. Furthermore, this optimization at inference time allows leveraging global scene information to inform our optimized part decompositions; in particular, we consider objects of the same predicted class with similar estimated geometry, and optimize them jointly, enabling more robust and scene-consistent part decompositions.

In summary, we present the following contributions:

- We propose to learn optimizable part-based latent priors for 3D shapes – Neural Part Priors, encoding part segmentation and part geometry into a latent space for the part types of each class category.
- Our learned, optimizable part priors enable test-time optimization over the latent spaces, enhanced with *inter-shape part-based constraints*, to fit partial, cluttered object geometry in real-world scanned scenes, resulting in robust and precise semantic part completion.
- We additionally propose a scene-consistent optimization, enhanced with *intra-shape constraints*, jointly optimizing over similar objects that provides globally-consistent part decompositions for repeated object instances in a scene.

2. Related Works

3D Object Detection and Instance Segmentation. 3D semantic scene understanding has seen rapid progress in recent years, with large-scale 3D datasets [4, 11, 19] and developments in 3D deep learning showing significant advances in object-level understanding of 3D scenes. Various methods explore learning on different 3D representations for object detection and instance segmentation, including volumetric grids [26, 49], point clouds [30, 38, 42, 43, 55], sparse voxel representations [18, 24], and hybrid approaches [26, 42]. These approaches have achieved impressive performance in detecting and segmenting objects in real-world observations of 3D scenes, but do not consider lower-level object part information that is requisite for many vision and robotics tasks, that involve object interaction and manipulation.

Recently, Bokhovkin et al. [3] proposed an approach to estimate part decompositions of objects in RGB-D scans, leveraging structural part type prediction and a pre-computed set of geometric part priors. Due to the use of dense volumetric part priors, the part reasoning is limited to coarse resolutions, and does not precisely match the input observed geometry. We also address the task of semantic part prediction and completion for objects in real-world 3D scans, but leverage a learned, structured latent space representing neural part priors, enabling part reasoning at high resolutions which optimizing to fit accurately to the observed scan geometry.

3D Scan Completion. As real-world 3D reconstructions are very often incomplete due to the complexity of the scene geometry and occlusions, various approaches have been developed to predict complete shape or scene geometry from partial observations. Earlier works focused on voxel-based scan completion for shapes [16, 54], with more recent works tackling the challenge of generating complete geometry from partial observations of large-scale scenes [12, 15, 17, 48], but without considering individual object instances. Several recent works propose to detect objects in an RGB-D scan and estimate the complete object geometries, leveraging voxel [3, 27] or point [38, 58] representations. Our approach to predicting part decompositions of objects inherently provides object completion as the predicted parts union; in contrast to previous approaches estimating object completion in RGB-D scans, we propose to characterize object parts as learned implicit priors, enabling test-time traversal of the latent space to fit accurately to observed input scan geometry.

Part Segmentation of 3D Shapes. Part segmentation for 3D shapes has been well-studied in shape analysis, typically focusing on understanding collections of synthetic shapes. Various methods have been developed for unsupervised part segmentation by finding a consistent segmentation across a set of shapes [22, 28, 29, 33, 46]. Recent deep learning based approaches have leveraged datasets of shapes with part anno-

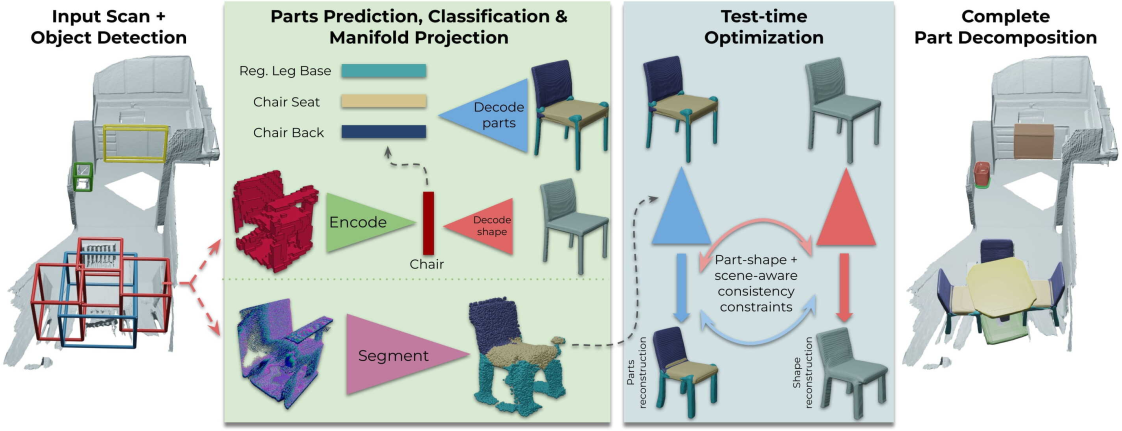


Figure 2. Method overview. From an input scan, we first detect 3D bounding boxes for objects. For each object, we predict their semantic part structure as a set of part labels and latent codes for each part. These latent codes map into the space of neural part priors, along with a full shape code used to regularize the shape structure. We then refine these codes at test time by optimizing to fit to the observed input geometry along with inter-object consistency between similar detected objects, producing effective part decompositions reflecting complete objects with scene consistency.

tations to learn part segmentation on new shapes [25, 31, 57]. In particular, approaches that learn part sequences and hierarchies to capture part structures have shown effective part segmentation for shapes [35, 36, 50, 51, 53, 56]. These approaches target single-object scenarios, whereas we construct a set of learned part priors that can be optimized to fit to real-world, noisy, incomplete scan geometry.

Neural Implicit Representations of 3D Shapes. Recently, we have seen significant advances in generative shape modeling with learned neural implicit representations that can represent continuous implicit surface representations, without ties to an explicit grid structure. Notably, DeepSDF [40] proposed an MLP-based network that predicts the SDF value for a given 3D location in space, conditioned on a latent shape code, which demonstrated effective modeling of 3D shapes while traversing the learned shape space. Such implicit representations have also been leveraged in hybrid approaches coupling explicit geometric locations with local implicit descriptions of geometry for shapes [20, 21] as well as scenes [41], without semantic meaning to the local decompositions. We propose to leverage the representation power of such learned continuous implicit surfaces to characterize semantic object parts that can be jointly optimized together to fit all parts of an object to a partial scan observation.

3. Method

3.1. Overview

We introduce Neural Part Priors (NPPs) to represent learned spaces of geometric object part priors, that enable joint part segmentation and completion of objects in real-world, incomplete RGB-D scans. From an input 3D scan $\mathcal{S}$, we first detect objects $\mathcal{O} = \{o_i\}$ in the scan characterized by their bounding boxes and orientations, then for each object, we predict its part decomposition into a part class categories (with corresponding part latent codes) and their corresponding complete geometry represented as signed distance fields (SDFs) and trained on part annotations for shapes. This enables holistic reasoning about each object in the scene and prediction of complete geometry in unobserved scan regions. Since captured real-world scene geometry contains significant incompleteness or noise, we model our geometric part priors based on complete, clean synthetic object parts, represented as a learned latent space over implicit part geometry functions. This enables test-time optimization over the latent space of parts to fit real geometry observations, enabling part-based object completion while precisely representing real object geometry. Rather than considering each object independently, we observe that repeated objects often occur in scenes under different partial observations, leading to inconsistent independent predictions; we thus jointly optimize across similar objects in a scene, where objects are considered similar if they share the same class category and predicted shape geometries are close by chamfer distance. This results in scene-consistent, high-fidelity characterizations of both object part semantics and complete object geometry. An overview of our approach is shown in Fig. 2.

3.2. Object Detection

From input 3D scan $\mathcal{S}$, we first detect objects in the scene, leveraging a state-of-the-art 3D object detection backbone from MLCVNet [55]. MLCVNet interprets $\mathcal{S}$ as a point cloud and proposes objects through voting [43] at multiple resolutions, providing an output set of axis-aligned bounding boxes for each detected object o_i. We extract the truncated signed distance field D_i for each o_i at 4mm resolution to use for test-time optimization. We then aim to characterize shape properties for o_i to be used for rotation estimation and test-time optimization, and interpret D_i as a 32^3 occupancy grid which is input to a 3D convolutional object encoder to produce the object's shape descriptor $\mathbf{s}_i \in \mathbb{R}^{256}$.

Initial Rotation Estimation. From $\mathbf{s}_i$, we use a 2-layer MLP to additionally predict an initial rotation estimate of the object as r_i^{init} around the up (gravity) vector of $\mathcal{S}$. We note that the up vectors of an RGB-D scan can be reliably estimated with IMU and/or floor estimation techniques [11]. The rotation estimation is treated as a classification problem across $n_r = 12$ bins of discretized angles ($\{0°, 30°, \ldots, 330°\}$), using a cross entropy loss. We use the estimated rotation r_i^{init} to resample D_i to approximate the canonical object orientation, from which we use to optimize for the final rotation r_i and the object part latent codes.

3.3. Learned Space of Neural Part Priors

We first learn a set of latent part spaces for each class category, where each part space represents all part types for the particular object category. To this end, we employ a function f_p characterized as an MLP to predict the implicit signed distance representation for each part geometry of the class category. In addition to the latent part space, we additionally train a proxy shape function f_s as an MLP that learns full shape geometry as implicit signed distances, which will serve as additional regularization during the part optimization. Both f_p and f_s are trained in auto-decoder fashion following DeepSDF [40]. Then each train shape part is embedded into a part latent space by optimizing for its code $\mathbf{z}_k^{\mathrm{p}} \in \mathbb{R}^{256}$ such that f_p conditioned on this code and the part type maps a positional encoding $\mathbf{x}^{pos} \in \mathbb{R}^{63}$ of a point $\mathbf{x} \in \mathbb{R}^3$ in the canonical space to SDF value d of the part geometry:

$$f_p : \mathbb{R}^{63} \times \mathbb{R}^{256} \times \mathbb{Z}_2^{N_c} \to \mathbb{R}, \qquad f_p(\mathbf{x}^{pos}, \mathbf{z}_k^{\mathrm{p}}, \mathbb{1}_{\mathrm{part}}) = d. \tag{1}$$

where $\mathbb{1}_{\mathrm{part}} \in \mathbb{Z}_2^{N_c}$ is a one-hot encoding of the part type for a maximum of N_c parts. Similar to NeRF [34], euclidean coordinates $\mathbf{x} \in \mathbb{R}^3$ are encoded using sin / cos functions with 10 frequencies $[2^0, ..., 2^9]$. The shape space is trained analogously for each class category where $\mathbf{z}_i^{\mathrm{s}} \in \mathbb{R}^{256}$ represents a shape latent code in the space:

$$f_s : \mathbb{R}^{63} \times \mathbb{R}^{256} \to \mathbb{R}, \qquad f_p(\mathbf{x}^{pos}, \mathbf{z}_i^{\mathrm{s}}) = d. \tag{2}$$

We train latent spaces of part and shape priors on the synthetic PartNet [36] dataset to characterize a space of complete parts and shapes, by minimizing the reconstruction error over all train shape parts, while optimizing for latent codes $\{\mathbf{z}_k^{\mathrm{p}}\}$ and weights of f_p. We use an ℓ_1 reconstruction loss with ℓ_2 regularization on the latent codes:

$$L = \sum_{j=1}^{N_p} |f_p(\mathbf{x}_j^{pos}, \mathbf{z}_k^{\mathrm{p}}, \mathbb{1}_{\mathrm{part}}) - D^{\mathrm{gt}}(\mathbf{x}_j^{pos})|_1 + \lambda ||\mathbf{z}_k^{\mathrm{p}}||_2^2 \tag{3}$$

for N_p points near the surface, the regularization weight λ =1e-5. We train f_s analogously.

3.4. Part Decompositions in Real Scenes

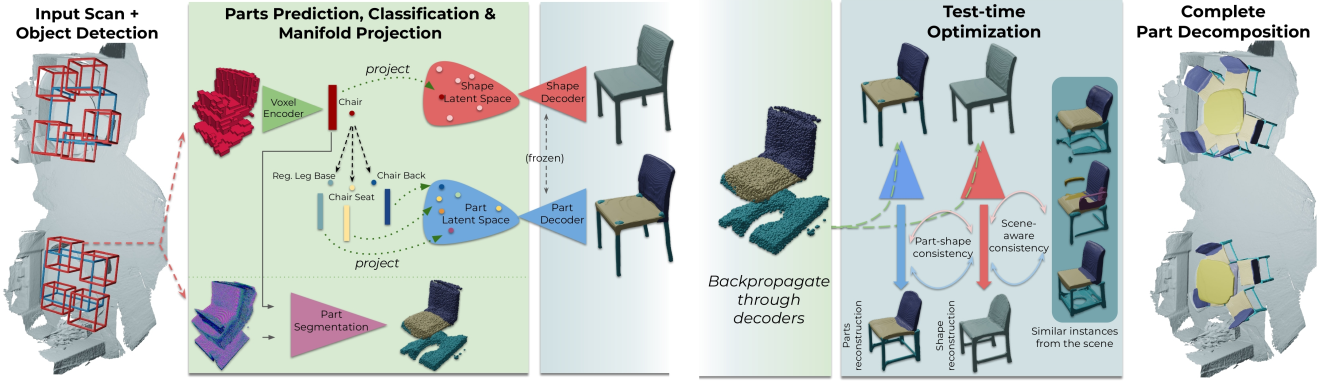
Once we have learned our latent space of parts, we can traverse them at test time to find the part-based decomposition of an object that best fits to its real-world observed geometry in a scene. Since real-world observations are typically incomplete, we can optimize for complete part decompositions based on strong priors given by the trained latent spaces. This allows for effective regularization by synthetic part characteristics (clean, complete) while fitting precisely to real observed geometry.

To guide this optimization for a detected object box o with its shape feature $\mathbf{s}$, we predict its high-level decomposition into a set of semantic part types $\{(c_k, \mathbf{p}_k)\}$, where $\mathbf{p}_k \in \mathbb{R}^{256}$ is a part feature descriptor and c_k the part class label. The i object suffix is discarded here for simplicity. The part optimization is initialized using $\{(c_k, \mathbf{p}_k)\}$.

To obtain the semantic part type predictions, we employ a message-passing graph neural network that infers part relations to predict the set of component part types. Similar to [35], from the shape feature $\mathbf{s}$ we use an MLP to predict at most $N_c = 10$ parts. For each potential part k, we predict its probability of existence, its part label c_k, its feature vector $\mathbf{p}_k$, and probabilities of physical adjacency (given by face connectivity between parts) between each pair of parts to learn structural part information. This produces the semantic description of the set of parts for the object $\{(c_k, \mathbf{p}_k)\}$, from the parts predicted with part existence probability > 0.5.

Projection to the Latent Part Space. We then learn a projection mapping from the part features $\{\mathbf{p}_k\}$ to the learned latent part space based on synthetic part priors, using a small MLP to predict $\{\tilde{\mathbf{z}}_k^{\mathrm{p}}\}$, as shown in Fig. 3(a). This helps to provide a close initial estimate in the latent part space in order to initialize optimization over these part codes to fit precisely to the observed object geometry. Similarly, we project the shape code $\mathbf{s}$ to the learned latent shape space with a small shape projection MLP to predict $\{\tilde{\mathbf{z}}^{\mathrm{s}}\}$, which we use to help regularize the part code optimization. Both of these projection MLPs are trained using MSE losses against the optimized train codes of the latent spaces.

Part Segmentation Estimation. In addition to our projection initialization, we estimate part segmentation $\{D^{\mathrm{p}}\}_{\mathrm{p}=1}^{N_{parts}}$ for the input object TSDF D over the full volume, representing part SDF geometry in the regions predicted as corresponding to the part p, where part segmentation regions cover the entire shape, including unobserved regions. This is used to guide part geometry predictions when optimizing at test time to fit to real observed input geometry. For each point $\mathbf{x} \in \mathbb{R}^3$ which has distance $< d_{trunc} = 0.16$m from the input object TSDF D, we classify it to one of the predicted parts $\{(c_k, \mathbf{p}_k)\}$ or background using a small PointNet-based [44] network. This



(a) Projection into part and shape spaces.

(b) Joint optimization to fit to input scan.

Figure 3. (a) Projection into the part and shape latent spaces along with part segmentation from input scan geometry. (b) Optimization at test time to fit to observed scan geometry while maintaining inter-part consistency within a shape and inter-shape consistency for geometrically similar objects.

segmentation prediction takes as input the corresponding shape feature $\mathbf{s}$, the initial estimated rotation r_i^{init}, and the 3D coordinates of $\mathbf{x}$, and is trained with a cross-entropy loss.

3.5. Joint Inter- and Intra-Shape Part Optimization

To obtain the final part decompositions, we traverse over the learned latent part space to fit to the observed input scan geometry, as shown in Fig. 3(b). From the initial estimated part codes $\{\tilde{\mathbf{z}}^{\text{p}}_k\}$ and shape code $\{\tilde{\mathbf{z}}^{\text{s}}\}$, their decoded part SDFs should match to each of $\{D^{\text{p}}\}_{\text{p}=1}^{N_{parts}}$. Since the part and shape latent spaces have been trained in the canonical shape space, we optimize for a refined rotation prediction r from r^{init} using iterative closest points [2, 45] between the sampled points near D and the initial shape estimate from projection $\tilde{\mathbf{z}}^{\text{s}}$. We use N_i sampled points near the observed input surface D (near being SDF values < 0.025m) for rotation refinement, with N the number of points not predicted as background during part segmentation.

While the predicted projected part and shape codes $\{\tilde{\mathbf{z}}^{\text{p}}_k\}, \{\tilde{\mathbf{z}}^{\text{s}}_k\}$ can provide a good initial estimate of the part decomposition of the complete shape, they represent synthetic part and shape priors that often do not fit the observed real input geometry. We thus optimize for part decompositions that best fit the input observations by minimizing the energy:

$$L = \sum_k L_{\text{part}} + L_{\text{shape}} + w_{\text{cons}} L_{\text{cons}}, \tag{4}$$

where L_{part} denotes the part reconstruction loss, L_{shape} a proxy shape reconstruction loss, L_{cons} a regularization to encourage global part consistency within the estimated shape, and w_{cons} is a consistency weight.

L_{part} is an ℓ_1 loss on part reconstruction:

$$L_{\text{part}} = \sum_{\text{p}=1}^{N_{parts}} \sum_{N^{\text{p}}} w_{trunc} |f_p(\mathbf{z}^{\text{p}}_k) - T_r(D^{\text{p}})| + \lambda ||\mathbf{z}^{\text{p}}_k||_2^2, \tag{5}$$

where N^{p} is the number of points classified to part p and w_{trunc} gives a fixed greater weight for near-surface points ($< d_{trunc} = 0.16$m).

L_{shape} is a proxy ℓ_1 loss on shape reconstruction:

$$L_{\text{shape}} = \sum_N w_{trunc} |f_s(\mathbf{z}^{\text{s}}) - T_r(D)| + \lambda ||\mathbf{z}^{\text{s}}||_2^2. \tag{6}$$

The regularization weight λ =1e-5 for both Eq. 5, 6. Finally, L_{cons} encourages all parts to reconstruct a shape similar to the optimized shape:

$$L_{\text{cons}} = \sum_N |f_p(\mathbf{z}^{\text{p}}_k) - f_s(\mathbf{z}^{\text{s}})|, \tag{7}$$

where $f_s(\mathbf{z}^{\text{s}})$ is frozen for L_{cons}. This allows for reconstructed parts to join together smoothly without boundary artifacts to holistically reconstruct a shape.

This produces a final optimized set of parts for each object in the scene, where parts both fit precisely to input geometry and represent the complete geometry of each part, even in unobserved regions. The final part geometries can be extracted from the SDFs with Marching Cubes [32] to obtain a surface mesh representation of the semantic part completion.

Scene-Consistent Optimization. Scenes often contain repeated instances of objects which are observed from different views, thus frequently resulting in inconsistent part decompositions when considered as independent objects. We propose a scene-consistent optimization between similar predicted objects, where objects in a scene are considered similar if their predicted class category is the same and the chamfer distance between their decoded shapes from $\tilde{\mathbf{z}}^{\text{s}}_k$ is $< \tau_s$.

For a set of N_{sim} similar objects in a scene, we collect together their predicted part segmentations and observed input SDF geometry in the canonical orientation based on $T_r(D^{\text{p}})$ to provide a holistic set of constraints across different partial observations to produce $\{D_i\}_{i=1}^{N_{sim}}$. The $\{D_i\}_{i=1}^{N_{sim}}$ are then aggregated to form $\mathbf{D}'$ by sampling a set of N_{avg} points near the surfaces of $\{D_i\}_{i=1}^{N_{sim}}$ where N_{avg} is the average number of points across the N_{sim} objects, and each point is assigned the minimum SDF value within its 30-point local neighbour-

	Chamfer Distance – Accuracy (↓)								Chamfer Distance – Completion (↓)							
Method	chair	table	cab.	bkshlf	bed	bin	class avg	inst avg	chair	table	cab.	bkshlf	bed	bin	class avg	inst avg
SG-NN [12] + MLCVNet [55] + PointGroup [30]	0.047	0.110	0.146	0.173	0.350	0.051	0.146	0.083	0.054	0.141	0.123	0.192	0.382	0.045	0.156	0.089
MLCVNet [55] + StructureNet [35]	0.024	0.074	0.104	0.166	0.424	0.039	0.138	0.061	0.028	0.129	0.118	0.154	0.352	0.037	0.136	0.067
Bokhovkin et al. [3]	0.029	0.073	0.099	0.168	0.244	0.036	0.108	0.056	0.031	0.095	**0.108**	**0.151**	0.236	0.038	0.110	0.059
Ours	**0.013**	**0.049**	**0.080**	**0.134**	**0.139**	**0.022**	**0.074**	**0.035**	**0.020**	**0.073**	0.110	0.193	**0.132**	**0.023**	**0.092**	**0.045**

Table 1. Evaluation of semantic part completion on Scan2CAD [1] in comparison to state-of-the-art part segmentation [30, 35] and semantic part completion [3]. Our optimizable part priors produce more accurate part decompositions.

hood (to help ensure that objects do not grow thicker in size from small-scale misalignments). Based on $\mathbf{D}'$, we then optimize for the part decompositions following Eq. 4.

3.6. Implementation Details

We first train our latent part and shape spaces on per-category on the synthetic PartNet [36] dataset. We then train the projection mapping into the learned part and shape spaces as well as the part segmentation. This is first pre-trained on synthetic PartNet data using virtually scanned incomplete inputs to take advantage of the large amount of synthetic data. To apply to real-world observations, we then fine-tune the projections and part segmentation on ScanNet [11] data using MLCVNet [55] detections on train scenes.

In test time, we optimize for part and shape codes using an Adam optimizer with learning rate of 3e-4 for 300 iterations and 3e-5 for the next 300 iterations. To enable more flexibility to capture input details, we optimize the decoder weights for parts and shape after 400 iterations. Optimization for each part takes ≈ 25 seconds. For further implementation and training details, we refer to the supplemental.

4. Results

We evaluate our Neural Part Priors for semantic part completion on real-world RGB-D scans from ScanNet [11]. We use the official train/val/test split of 1045/156/ 312 scans. In order to evaluate part segmentation in these real-world scenes, we use the Scan2CAD [1] annotations of CAD model alignments from ShapeNet [5] to these scenes, and the PartNet [36] part annotations for the ShapeNet objects. To construct our part latent space, we train on PartNet and train a projection from train ScanNet objects to their PartNet annotations; we also train all baselines on the same ScanNet+PartNet data. We consider 6 major object class categories representing the majority of parts, comprising a total of 28 part types. All methods are provided the same MLCVNet bounding boxes, which contain at least 40% of the closest annotated shape from Scan2CAD [1] dataset.

Evaluation Metrics. Since PartNet shapes aligned to real-world ScanNet geometry do not have precise geometric alignments, due to inexact synthetic-real associations, our evaluation metrics aim to characterize the quality of part decompositions that fit well to the real-world geometry, as well as accurately represent complete object geometry. We evaluate *accuracy* of the part segmentation with respect to the ScanNet object, and object *completion* with respect to the complete object geometry from PartNet.

Accuracy measures part segmentation predictions with respect to PartNet labels projected onto ScanNet object geometry, as a single-sided chamfer distance from the partial ScanNet object to the predicted part decomposition (as we lack complete real-world geometry available for evaluation in the other direction). *Completion* evaluates a bi-directional Chamfer distance between the predicted part decomposition against the PartNet labeled shape.

For evaluation, we sample $10,000$ points per part from predicted and ground-truth mesh surfaces and transform them to the ScanNet coordinate space. Each shape instance is evaluated over a union of predicted and ground-truth parts and then summed to represent shape evaluation. All shape evaluations are averaged to form instance average and the mean of shape category averages results in class average.

To evaluate part segmentation of only the observed input geometry without considering completion, we use Chamfer distance and IoU. We project part labels from aligned PartNet shapes and predicted mesh parts onto the ScanNet mesh surface to obtain ground truth and predicted part segmentations. For further details, we refer to the supplemental.

Comparison to state of the art. Tab. 1 shows a comparison to state of the art on semantic part completion for real-world ScanNet scans, with qualitative results shown in Fig. 4. We compare with Bokhovkin et al. [3], which leverages coarse 32^3 pre-computed geometric priors for semantic part completion, the state-of-the-art part segmentation approaches Struc-

	Chamfer Distance (↓)								IoU (↑)							
Method	chair	table	cab.	bkshlf	bed	bin	class avg	inst avg	chair	table	cab.	bkshlf	bed	bin	class avg	inst avg
SG-NN [12] + MLCVNet [55] + PointGroup [30]	0.056	0.121	0.110	0.161	0.406	0.034	0.148	0.088	0.257	0.390	**0.300**	0.229	0.306	**0.488**	0.328	0.293
MLCVNet [55] + StructureNet [35]	0.025	0.101	0.092	**0.090**	0.359	0.041	0.118	0.059	0.480	0.356	0.195	0.331	0.267	0.342	0.329	0.413
Bokhovkin et al. [3]	0.027	0.059	0.095	0.102	0.207	0.031	0.087	0.049	**0.548**	**0.542**	0.224	0.328	0.442	0.375	0.409	**0.490**
Ours	**0.021**	**0.051**	**0.076**	0.094	**0.141**	**0.025**	**0.068**	**0.038**	0.489	**0.542**	0.272	**0.386**	**0.476**	0.340	**0.416**	0.461

Table 2. Evaluation of part segmentation on Scan2CAD [1]. We evaluate part segmentation only on observed scan geometry, in comparison with state-of-the-art part segmentation [30, 35] and semantic part completion [3].

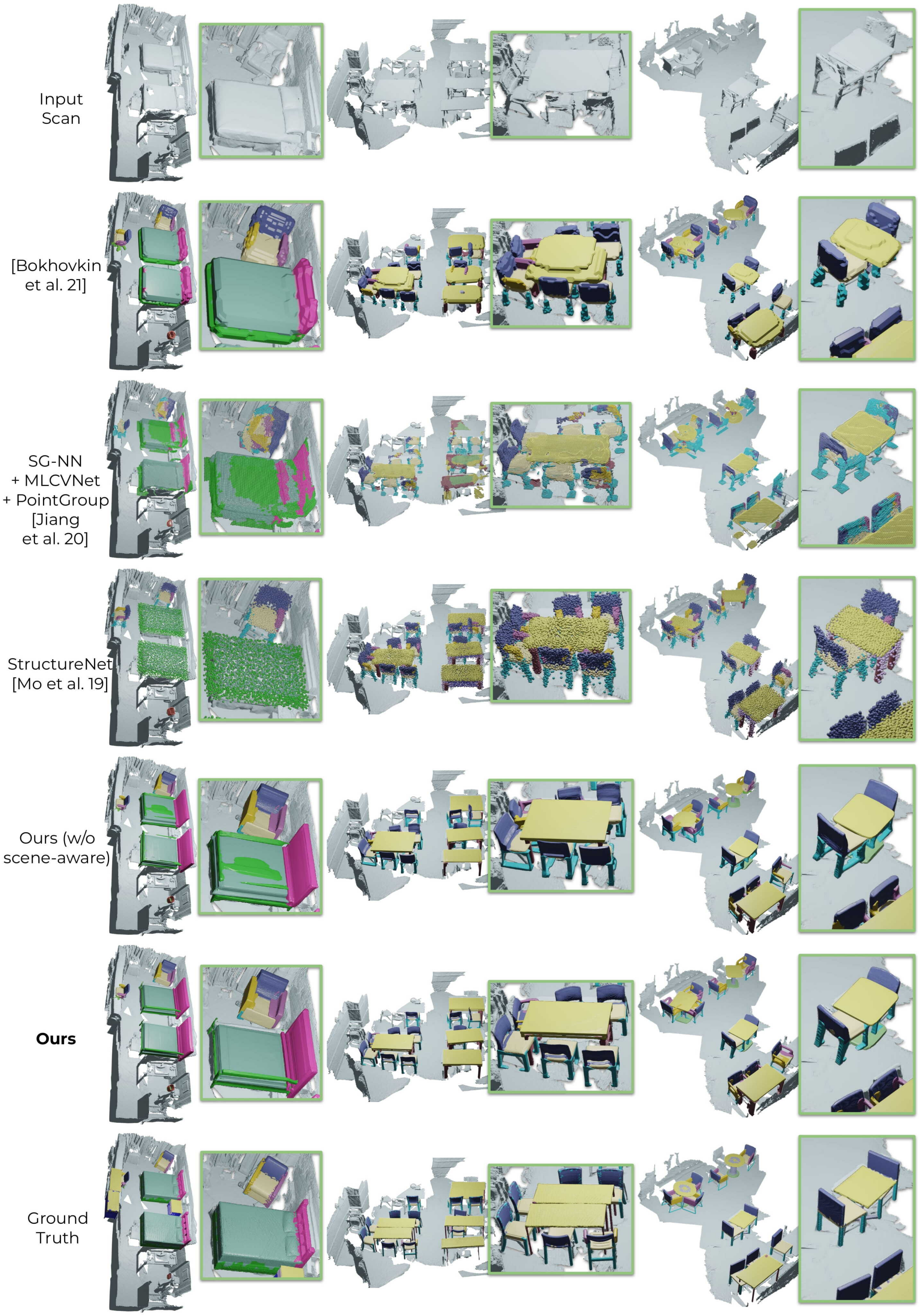

Input Scan
[Bokhovkin et al. 21]
SG-NN + MLCVNet + PointGroup [Jiang et al. 20]
StructureNet [Mo et al. 19]
Ours (w/o scene-aware)
Ours
Ground Truth

Figure 4. Qualitative comparison of NPPs with point [30, 35] and voxel-based [3] state of the art on ScanNet scans with Scan2CAD+PartNet ground truth. Our joint optimization across part priors enables more consistent, accurate part decompositions.

tureNet [35] and PointGroup [30]. Each of these methods uses the same object detection results from MLCVNet [55]; since PointGroup does not predict any geometric completion, we provide additional scan completion from SG-NN [12]. For Bokhovkin et al. [3] we apply Marching Cubes to voxels to extract a surface with which to evaluate. In contrast to these approaches which estimate part decompositions directly and independently, our NPPs enable joint optimization over parts to fit precisely to the observed scan geometry, resulting in improved accuracy and completion performance. Additional qualitative results are shown in the supplemental.

Part segmentation on 3D scans. We also evaluate part segmentation in Tab. 2, which considers segmentation of only the observed scan geometry, without any geometric completion. We intersect each method's part predictions with the original scan geometry to evaluate this part segmentation, in comparison with Bokhovkin et al. [3], StructureNet [35], and PointGroup [30]. By jointly optimizing over the parts of an object to fit to its observed scan geometry, our approach improves notably in part segmentation performance.

What is the effect of scene-consistent optimization? Tab. 3 and Fig. 4 show the effect of scene-consistent optimization. This improves results over *w/o Scene Consistency*, with a stronger effect on categories that more often have repeated instances (e.g., chair, table, bed in offices, classrooms, hotel rooms). This more holistic optimization enables more consistent reasoning about objects and their parts in a scene.

What is the impact of test-time optimization and projection initialization? We consider our approach without using a learned projection mapping to the latent part space as initialization for test-time optimization to fit the observed scan geometry and also evaluate only the projection mapping without test-time optimization, shown in Tab. 3. The initial projection helps significantly to obtain a good initialization for test-time optimization (*w/o Projection Map*), while the projection results provide a good estimate but imprecise fit to the observed scene geometry (*w/o Test-Time Opt*).

What is the effect of segmentation and full-shape constraints? The effects of full-shape constraints (*w/o Full-shape Constr.*) and dense segmentation (*w/o Segmentation*) are evaluated in Tab. 3. The full-shape constraint helps to maintain the global shape consistency of the optimized parts during TTO. Dense segmentation guides the TTO optimization constraints (e.g., avoids fitting to clutter; prevents self-intersections between optimized reconstructed parts).

Limitations. While our NPPs shows strong promise towards accurate, high resolution characterization of semantic parts in real-world scenes required for finer-grained semantic scene understanding, various limitations remain. Our part latent space is trained in its canonically oriented space, and while we can optimize for shape orientations with ICP, a joint optimization or an equivariant formulation can potentially resolve sensitivity to misaligned orientations. Finally, our scene-consistent optimization for similar shapes in a scene makes an important step towards holistic scene reasoning, but does not consider higher-level, stylistic similarity that is often shared across objects in the same scene (e.g., matching furniture set for desk, shelves, chair) which could provide notable insight towards comprehensive scene understanding.

5. Conclusion

We have presented Neural Part Priors, which introduces learned, optimizable part priors for fitting complete part decompositions to objects detected in real-world RGB-D scans. We learn a latent part space over all object parts, characterized with learned neural implicit functions. This allows for traversing over the part space at test time to jointly optimize across all parts of an object such that it fits to the observed scan geometry while maintaining consistency with any similar detected objects in the scan. This results in improved part segmentation as well as completion in noisy, incomplete real-world RGB-D scans. We hope that this help to open up further avenues towards holistic part-based reasoning in real-world environments.

Acknowledgements. This work was supported by the Bavarian State Ministry of Science and the Arts coordinated by the Bavarian Research Institute for Digital Transformation (bidt) and the German Research Foundation (DFG) Grant "Learning How to Interact with Scenes through Part-Based Understanding."

	Chamfer Distance (↓) – Accuracy								Chamfer Distance (↓) – Completion							
Method	chair	table	cab.	bkshlf	bed	bin	class avg	inst avg	chair	table	cab.	bkshlf	bed	bin	class avg	inst avg
w/o Projection Map.	0.032	0.156	0.140	0.249	0.352	0.029	0.160	0.081	0.026	0.144	0.137	0.192	0.309	0.025	0.139	0.070
w/o Synthetic Pretrain	0.026	0.079	0.116	0.158	0.152	0.035	0.094	0.052	0.029	0.097	0.132	0.206	0.234	0.033	0.122	0.061
w/o Test-Time Opt.	0.019	0.060	0.105	0.145	0.200	0.023	0.093	0.047	0.022	0.085	0.127	0.183	0.198	0.024	0.107	0.052
w/o Full-shape Constr.	0.016	0.051	0.090	0.151	0.168	0.025	0.083	0.040	0.023	0.075	0.121	0.202	0.157	0.026	0.101	0.050
w/o Segmentation	**0.013**	**0.046**	**0.065**	**0.120**	**0.062**	0.023	**0.058**	**0.030**	0.035	0.084	0.118	0.195	0.203	0.034	0.112	0.061
w/o Scene Consistency	0.016	0.053	0.080	0.139	0.152	0.023	0.077	0.038	**0.020**	0.074	**0.108**	**0.180**	0.150	0.025	0.093	**0.045**
Ours	**0.013**	0.049	0.080	0.134	0.139	**0.022**	0.074	0.035	**0.020**	**0.073**	0.110	0.193	**0.132**	**0.023**	**0.092**	**0.045**

Table 3. Ablation study evaluating semantic part completion on Scan2CAD [1]. We show the effect of our projection mapping initialization, and test-time optimization to fit to the for our design decisions. Overall, using both results in the best performance in RGB-D scan part decompositions. In bold – the best result, underlined – the second best result.

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 *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2614–2623, 2019. 6, 8, 12, 15, 17, 18

[2] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. Spie, 1992. 5

[3] Alexey Bokhovkin, Vladislav Ishimtsev, Emil Bogomolov, Denis Zorin, Alexey Artemov, Evgeny Burnaev, and Angela Dai. Towards part-based understanding of rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7484–7494, 2021. 1, 2, 6, 7, 8, 12, 13, 14, 15

[4] Angel X. Chang, Angela Dai, Thomas A. Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. In *2017 International Conference on 3D Vision, 3DV 2017, Qingdao, China, October 10-12, 2017*, pages 667–676, 2017. 1, 2

[5] 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. 6

[6] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5932–5941, 2019. 1

[7] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. pages 6968–6979, 06 2020. 1

[8] Julian Chibane, Aymen Mir, and Gerard Pons-Moll. Neural unsigned distance fields for implicit function learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, December 2020. 1

[9] Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. Robust reconstruction of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5556–5565, 2015. 1

[10] Christopher B. Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3075–3084, 2019. 1

[11] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2, 4, 6, 15, 17, 19

[12] Angela Dai, Christian Diller, and Matthias Nießner. Sg-nn: Sparse generative neural networks for self-supervised scene completion of rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 849–858, 2020. 2, 6, 8, 12, 15

[13] Angela Dai and Matthias Nießner. 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–468, 2018. 1

[14] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics (ToG)*, 36(4):1, 2017. 1

[15] 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

[16] 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. 2

[17] Angela Dai, Yawar Siddiqui, Justus Thies, Julien Valentin, and Matthias Nießner. Spsg: Self-supervised photometric scene generation from rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1747–1756, 2021. 2

[18] Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 3d-mpa: Multi-proposal aggregation for 3d semantic instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9031–9040, 2020. 1, 2

[19] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 2

[20] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3d shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4857–4866, 2020. 3

[21] Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T Freeman, and Thomas Funkhouser. Learning shape templates with structured implicit functions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7154–7164, 2019. 3

[22] Aleksey Golovinskiy and Thomas Funkhouser. Consistent segmentation of 3d models. *Computers & Graphics*, 33(3):262–269, 2009. 2

[23] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 9224–9232, 2018. 1

[24] Lei Han, Tian Zheng, Lan Xu, and Lu Fang. Occuseg: Occupancy-aware 3d instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2940–2949, 2020. 1, 2

[25] 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):1–12, 2019. 3

[26] Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4421–4430, 2019. 1, 2

[27] Ji Hou, Angela Dai, and Matthias Nießner. Revealnet: Seeing behind objects in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2098–2107, 2020. 2

[28] Ruizhen Hu, Lubin Fan, and Ligang Liu. Co-segmentation of 3d shapes via subspace clustering. In *Computer graphics forum*, volume 31, pages 1703–1713. Wiley Online Library, 2012. 2

[29] Qixing Huang, Vladlen Koltun, and Leonidas Guibas. Joint shape segmentation with linear programming. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, pages 1–12, 2011. 2

[30] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4867–4876, 2020. 2, 6, 7, 8, 12, 13, 14, 15

[31] Evangelos Kalogerakis, Melinos Averkiou, Subhransu Maji, and Siddhartha Chaudhuri. 3d shape segmentation with projective convolutional networks. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3779–3788, 2017. 3

[32] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 5

[33] Tiange Luo, Kaichun Mo, Zhiao Huang, Jiarui Xu, Siyu Hu, Liwei Wang, and Hao Su. Learning to group: A bottom-up framework for 3d part discovery in unseen categories. *arXiv preprint arXiv:2002.06478*, 2020. 2

[34] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 4

[35] Kaichun Mo, Paul Guerrero, Li Yi, Hao Su, Peter Wonka, Niloy Mitra, and Leonidas J Guibas. Structurenet: Hierarchical graph networks for 3d shape generation. *arXiv preprint arXiv:1908.00575*, 2019. 3, 4, 6, 7, 8, 12, 13, 14, 15

[36] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 909–918, 2019. 3, 4, 6, 15, 16, 17

[37] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, pages 127–136. IEEE, 2011. 1

[38] Yinyu Nie, Ji Hou, Xiaoguang Han, and Matthias Nießner. Rfd-net: Point scene understanding by semantic instance reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4608–4618, 2021. 2

[39] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (ToG)*, 32(6):1–11, 2013. 1

[40] 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, 3, 4

[41] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 523–540. Springer, 2020. 3

[42] Charles R Qi, Xinlei Chen, Or Litany, and Leonidas J Guibas. Imvotenet: Boosting 3d object detection in point clouds with image votes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4404–4413, 2020. 2

[43] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9277–9286, 2019. 2, 3

[44] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 77–85, 2017. 4, 18, 20

[45] Szymon Rusinkiewicz and Marc Levoy. Efficient variants of the icp algorithm. In *Proceedings third international conference on 3-D digital imaging and modeling*, pages 145–152. IEEE, 2001. 5

[46] Oana Sidi, Oliver van Kaick, Yanir Kleiman, Hao Zhang, and Daniel Cohen-Or. Unsupervised co-segmentation of a set of shapes via descriptor-space spectral clustering. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, pages 1–10, 2011. 2

[47] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Proc. NeurIPS*, 2020. 1

[48] 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. 2

[49] Hsiao-Yu Fish Tung, Ricson Cheng, and Katerina Fragkiadaki. Learning spatial common sense with geometry-aware recurrent networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2595–2603, 2019. 2

[50] Oliver Van Kaick, Kai Xu, Hao Zhang, Yanzhen Wang, Shuyang Sun, Ariel Shamir, and Daniel Cohen-Or. Co-

hierarchical analysis of shape structures. *ACM Transactions on Graphics (TOG)*, 32(4):1–10, 2013. 3

[51] Yanzhen Wang, Kai Xu, Jun Li, Hao Zhang, Ariel Shamir, Ligang Liu, Zhiquan Cheng, and Yueshan Xiong. Symmetry hierarchy of man-made objects. In *Computer graphics forum*, volume 30, pages 287–296. Wiley Online Library, 2011. 3

[52] Thomas Whelan, Stefan Leutenegger, R Salas-Moreno, Ben Glocker, and Andrew Davison. Elasticfusion: Dense slam without a pose graph. Robotics: Science and Systems, 2015. 1

[53] Rundi Wu, Yixin Zhuang, Kai Xu, Hao Zhang, and Baoquan Chen. Pq-net: A generative part seq2seq network for 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 829–838, 2020. 3

[54] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1912–1920, 2015. 2

[55] Qian Xie, Yu-Kun Lai, Jing Wu, Zhoutao Wang, Yiming Zhang, Kai Xu, and Jun Wang. Mlcvnet: Multi-level context votenet for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10447–10456, 2020. 2, 3, 6, 8, 12, 15, 19

[56] Li Yi, Leonidas Guibas, Aaron Hertzmann, Vladimir G Kim, Hao Su, and Ersin Yumer. Learning hierarchical shape segmentation and labeling from online repositories. *arXiv preprint arXiv:1705.01661*, 2017. 3

[57] Li Yi, Vladimir G Kim, Duygu Ceylan, I-Chao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas Guibas. A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics (ToG)*, 35(6):1–12, 2016. 3

[58] Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J Guibas. Gspn: Generative shape proposal network for 3d instance segmentation in point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3947–3956, 2019. 2