LASA: Instance Reconstruction from Real Scans using A Large-scale Aligned Shape Annotation Dataset

Haolin Liu[1,2*], Chongjie Ye[1,2*], Yinyu Nie[3], Yingfan He[1,2], Xiaoguang Han[2,1†]

[*]equal contribution [†]corresponding author

[1]FNii, CUHKSZ [2]SSE, CUHKSZ [3]Technical University of Munich

gap-lab-cuhk-sz.github.io/LASA

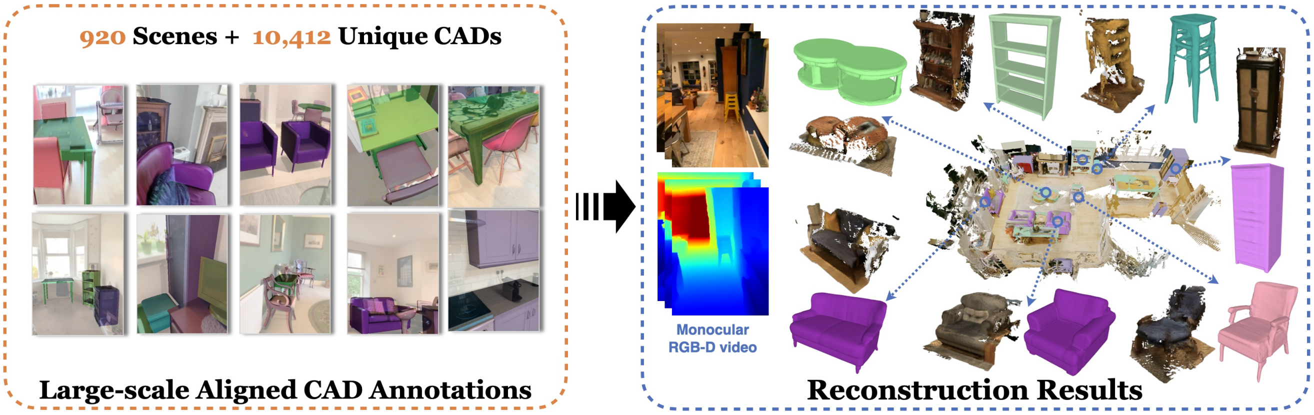
920 Scenes + 10,412 Unique CADs
Large-scale Aligned CAD Annotations
Monocular RGB-D video
Reconstruction Results

Figure 1. We introduce LASA, a Large-scale Aligned Shape Annotation Dataset containing 10,412 unique object CAD models aligned with 920 real-world scene scans. Supported by LASA, we achieve state-of-the-art in real-world object reconstruction and 3D object detection.

Abstract

Instance shape reconstruction from a 3D scene involves recovering the full geometries of multiple objects at the semantic instance level. Many methods leverage data-driven learning due to the intricacies of scene complexity and significant indoor occlusions. Training these methods often requires a large-scale, high-quality dataset with aligned and paired shape annotations with real-world scans. Existing datasets are either synthetic or misaligned, restricting the performance of data-driven methods on real data. To this end, we introduce LASA, a Large-scale Aligned Shape Annotation Dataset comprising 10,412 high-quality CAD annotations aligned with 920 real-world scene scans from ArkitScenes, created manually by professional artists. On this top, we propose a novel Diffusion-based Cross-Modal Shape Reconstruction (DisCo) method. It is empowered by a hybrid feature aggregation design to fuse multi-modal inputs and recover high-fidelity object geometries (see Fig. 1). Besides, we present an Occupancy-Guided 3D Object Detection (OccGOD) method and demonstrate that our shape annotations provide scene occupancy clues that can further improve 3D object detection. Supported by LASA, extensive experiments show that our methods achieve state-of-the-art performance in both instance-level scene reconstruction and 3D object detection tasks.

1. Introduction

The widespread use of hand-held RGB-D sensors has facilitated the effortless acquisition of indoor scene scans. However, these scans often suffer from noises and incompleteness due to limitations in sensor accuracy, the complexity of indoor environments, and occlusion among objects. This further limits its applications in scenarios such as VR/AR and 3D industry where a complete and high-quality reconstruction is desired. This shortage absorbs great attention in 3D vision and graphics community, particularly in the realm of indoor instance-level scene reconstruction. In this task, the objective is to reconstruct the shapes of observed objects based on 3D scans or images captured by sensors. Many advances [8, 36, 56, 58, 61] have been seen by leveraging the power of deep learning methods for this task. They are data-driven and demand a substantial number of paired scene scans and objects' CAD ground truths.

Existing data-driven methods rely on two types of datasets. [1, 7–9, 19, 27, 39, 50, 52, 55–58] utilized synthetic datasets [5, 10, 15] by synthesizing images and scans that mimic real-world distributions. Synthetic data provides CAD models perfectly aligned with input observations (images or scans) though. Models trained on it are vulnerable to domain gaps from the real world [53] that could impair the generalizability. Scan2CAD[2] attempts to bridge this gap by annotating objects' CAD in real-world scene scans. However, their CADs are retrieved from a synthetic database [5] and their poses are manually adjusted to roughly align the object scans. However, retrieved shapes are often unlike the real observed objects, introducing shape misalignment issues. Many works [13, 18, 28, 32, 47, 48] using it as instance shape supervision are potentially biased with inferior alignment. In summary, The absence of a well-aligned real-world dataset barriers the further advancement of the community.

To surmount this challenge, we present a new dataset **LASA**, a **L**arge-scale **A**ligned **S**hape **A**nnotation dataset that contains 10,412 high-quality instance CAD annotations **meticulously crafted** by skilled artists. Each CAD shape is designed and placed to precisely align with objects' scans from 920 real-world scene scans in ArkitScene [3]. We deliberately annotate objects in ArkitScene instead of ScanNet[11] because it provides scans obtained from both accurate Laser sensors and hand-held RGB-D sensors. The accurate Laser scan benefits high-quality and aligned manual annotations. Meanwhile, the less accurate scans from hand-held devices support research on reconstruction and completion using consumer-level devices.

A large-scale scan dataset with high-quality instance CAD annotations empowers many downstream applications. We first investigate how LASA benefits indoor instance-level scene reconstruction. Given an indoor scene, images and scene scans can be obtained through scanning. Typically, the initial step involves 3D object detection, after which the partial point clouds and multi-view images of each detected instance are acquired, serving as visual cues for subsequent shape reconstruction. Point cloud provides natural 3D information though, they are often noisy and incomplete. On the other hand, images present rich appearance clues but lack 3D constraints. Inspired by the complementary nature of these two modalities in describing object surfaces, we advocate utilizing both modalities as inputs to fuse their advantages. Recently, diffusion models have shown appealing performance in shape generation [7, 19, 44, 61, 64]. We advance it further for instance reconstruction and propose Diffusion-based Cross-modal Shape Reconstruction, namely **DisCo**. DisCo is a triplane diffusion model tailored to accommodate multi-modal inputs, towards robust real-world object reconstruction, where a hybrid feature aggregation layer is proposed to aggregate and align two input modalities effectively. Supported by the LASA dataset, extensive experiments show that our method achieves state-of-the-art reconstruction performance with real-world inputs.

Moreover, the LASA dataset has full annotations covering every instance within each of the 920 real-world scenes. This extensive coverage enables LASA to provide scene-level occupancy labels. We further explore how they can affect indoor scene 3D object detection. 3D Object detection [37, 42, 49] usually comprises a backbone and a detection head. We propose to integrate an occupancy prediction head for occupancy-guided 3D object detection, namely **OccGOD**. Our experiments demonstrate that incorporating occupancy prediction leads to substantial improvements in detection performance. In summary, our key contributions are four-fold:

- We introduce a large-scale dataset, LASA. It contains 10,412 manually crafted, high-quality instance CAD annotations geometrically aligned with 920 real-world scene scans.
- We propose DisCo, a novel diffusion-based method that leverages hybrid representation, effectively interacting with both input partial point cloud and multi-view images, achieving state-of-the-art reconstruction.
- With LASA's scene-level annotation, we introduce occupancy-guided 3D object detection (OccGOD) with decent improvements.
- We strongly believe the large-scale dataset of well-aligned shape annotations can break the bottleneck of current research on 3D indoor scene understanding and reconstruction.

2. Related Works

2.1. Indoor Instance Shape Dataset

Recently, many advances have been seen in learning-based 3D object and scene reconstruction from images and videos. However, numerous challenges persist due to the limitations of existing datasets.

These learning-based approaches are usually trained either on existing synthetic datasets or real-world datasets. While synthetic datasets are demonstrated valuable for training models, they lack realism. Large synthetic object collections like ShapeNet[5] and ABO[10] provide diverse 3D models but lack environment context. Synthetic scene datasets such as 3D-Front[14], Replica[45], and Structured3D[63] possess complete synthetic environments with CAD annotations for objects. However, models trained on these datasets often struggle to generalize with real inputs due to substantial domain gaps[53].

Some works [2, 29, 34, 46, 53, 54] annotate CAD

Dataset	Aligned	#Scenes	#CADs	Sensor Type	Annotation Method
Scan2CAD [2]	-	1,506	3,049	RGB-D	Retrieval
CAD-Estate [29]	-	19,512	12,024	RGB	Retrieval
IKEA [16]	✓	null	90	RGB	Retrieval
Pix3D [46]	✓	null	219	RGB	Retrieval
ScanSalon [53]	✓	413	800	RGB-D	Artist
Aria's Digital Twin [34]	✓	2	370	RGB-D	Artist
LASA (Ours)	✓	920	10,412	RGB-D	Artist

Table 1. Comparisons with existing 3D indoor datasets with instance shape annotations

models on real-world data to bridge the domain gaps. Scan2CAD [2] and CAD-estate [29] have provided scanned object-CAD pairs by retrieving them from ShapeNet [5] and are further manually aligned to the real-world scenes. However, these retrieved CAD models lack alignment with real objects, potentially biasing data-driven reconstruction methods towards inferior reconstruction. This motivates us to build a real-world dataset with well-aligned CAD annotations. Aria's Digital Twin [34] and ScanSalon [53] provide such CADs though, the limited quantities restrict their application for data-hungry tasks. Datasets such as Pix3d [46] and PASCAL3D+ [54] supply single-view images with aligned shapes but are limited in scene modalities lacking point cloud and multi-view images.

2.2. 3D Shape Reconstruction

Object-level Reconstruction Existing approaches leverage images or partial point clouds as input. Images based reconstruction methods accept either single-view [6, 17, 19, 20, 25, 27, 30, 31, 33, 35, 50, 55, 57, 62, 64] or sparse multi-view images [4, 9, 39, 56, 59] as inputs for shape reconstruction. They extract 2D image features and utilize them for shape reconstruction. Some [27, 43] leverage 2D pixel-aligned local features for reconstruction, demonstrating appealing performance for high-quality reconstruction. Other works [1, 8, 12, 26, 40, 52, 58, 61] conduct shape reconstruction from noisy, incomplete partial point clouds. Both paradigms have their advantages, with image-based methods perceiving rich 2D appearance features, while point-based methods process native 3D signals. [21] proposed to combine both inputs to fuse their advantages. Our proposed approach also accepts both inputs, focusing on effectively aggregating and aligning local features from both input sources. Since diffusion model [7, 19, 44, 51, 61, 64] have demonstrated strong capabilities in both shape generation and reconstruction, we opt for a diffusion model for robust and high-quality reconstruction.

Instance-level Scene Reconstruction Recent instance scene reconstruction methods [13, 18, 21–23, 28, 32, 47, 48] follow a detect-then-reconstruct pipeline from a single scan. First, these methods use a 3D object detector to localize objects in either scene scans or videos. Subsequently, the detected 3D objects are fed into a 3D shape reconstruction module individually to obtain object shapes. Finally, the reconstructed 3D object shapes are placed back at their original locations to obtain a reconstruction of the full scene. These methods commonly utilize the Scan2CAD [2] dataset for training, where the inferior alignment limits their reconstruction quality. In contrast, the CAD models in our LASA dataset are manually crafted by artists to guarantee alignment, which we hope can lay a foundation for future research in this area.

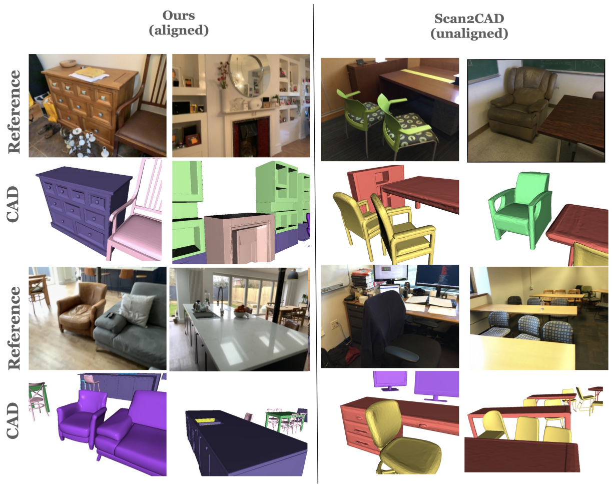

Figure 2. Visual comparison between aligned and unaligned CAD annotations

3. LASA Dataset

LASA is a large-scale dataset that contains 10,412 unique CAD models covering 920 scenes across 17 categories. Rather than relying on a pre-existing CAD database to annotate scenes, LASA engages professional artists to manually create aligned CAD models with 3D scans. Our annotations provide precise and consistent (as shown in Fig. 2) training data for data-driven reconstruction algorithms.

3.1. Data Annotation

LASA is built upon the 3D laser scans from ArkitScenes since it is accurate with high resolution, which is critical for high-fidelity CAD annotations.

Data Preprocessing. Each laser scan from ArkitScenes is exceptionally dense with 1GB+ data storage. To improve annotation efficiency, we downsample the scans to a 4mm density. Since ArkitScenes does not publicly provide alignment transformations between the laser scans and the RGB-D scans, we utilize a coarse-to-fine registration method to calculate the transformation matrix (see our supplemental for details). With the transformation matrix, we align the laser data with the RGB-D sequence coordinates from ArkitScenes. We then use ArkitScenes' 3D bounding box annotations to partition the aligned laser point clouds for each single object. These segmented point clouds are transformed into the canonical space. Furthermore, for each object, we select 2-5 frames from RGB-D scans that maximize the 2D projection area of its 3D bounding box. These selected frames serve as references for annotation.

Shape Annotation. The CAD annotation process involves a team of 35 artists working over 4 months. With preprocessed point clouds and reference images, each artist spent approximately 69 minutes designing a single model. Each model is annotated with Autodesk Maya or Cinema 4D.

Shape Verification. We involve a shape verification procedure for annotation quality control. This procedure has both algorithmic validation and manual reviews to thoroughly evaluate CAD model's accuracy against ground truth scans and images. Our multi-step verification process includes

- Senior Review: After initial annotations, 6 senior designers reviewed every CAD model to verify quality, accuracy, and reliability by manually cross-checking against the 3D scans. Any models that failed to meet the standards were flagged for rework.
- Geometry Alignment: We matched the CAD models to aligned laser scans and calculated the unidirectional Chamfer distance between them. This quantified the raw geometric alignment error for the CAD surface compared to the ground truth scan.
- View Alignment: We also verified alignment in the pixel level, where we rendered 112,639 images across all scenes by positioning the CAD models in the view frustum of RGB-D sensors. Crowd workers performed a manual inspection to check if the rendered views overlay on the real images. They checked for inconsistencies along object edges and intricate details which would indicate misalignment. This pixel-level evaluation ensured precise alignment. Any rendering mismatches were fixed by re-annotating the CAD model.

3.2. Dataset Statistics

Tab. 1 shows the statistics of LASA compared to the existing datasets. LASA contains a comparable number of unique CAD models to CAD-Estate [29], which was previously the largest scene CAD dataset. However, LASA provides better CAD annotation quality against their retrieved CAD models. Additionally, LASA demonstrates greater shape diversity than Scan2CAD [2], with over 3 times as many unique CAD models.

Among all aligned datasets, LASA stands out with a total of 10,412 CADs. This is 13 times more than ScanSalon [53] and 28 times more than Aria's Digital Twin dataset[34]. Unlike IKEA [16] and Pix3D [46] which are annotated on single-view RGB images, LASA captures full RGB-D sequences. This enables a wider range of downstream applications compared to static image datasets.

Furthermore, we compare LASA with Scan2CAD in terms of alignment quality by measuring the Chamfer distance between CAD annotation and the scene scans, which are **0.161** vs **0.269**. LASA provides much better aligned data.

4. Instance-level Scene reconstruction

We propose a Diffusion-based Cross-modal Shape Reconstruction method (DisCo). DisCo is a diffusion-based model to pursue high-fidelity 3D shape reconstruction from partial point clouds and multi-view images. Various representation including latent sets [61], volumetric grids [7, 64], and triplanes [19, 39, 51] are popular choice for diffusion model. We opt for the triplane since its efficiency enables higher output resolution compared to volumetric grids; while reserving 3D structure compared to latent sets. We employ latent triplane diffusion, where a triplane variational auto-encoder (VAE) first encodes shape into triplane latent space (in Sec. 4.1). Subsequently, a triplane diffusion model operates on this latent space for 3D shape reconstruction conditioned on both partial points and multi-view images (in Sec. 4.2). The overall pipeline of DisCo is shown in the upper part of Fig. 3.

It has been demonstrated that using multiple input modalities (images and scans) presents complementary benefits for semantic scene completion [21]. In DisCo, we also fuse image and point features and introduce **Hybrid Feature Aggregation Layer** (in Sec. 4.3). In this layer, we utilize a hybrid representation combining a triplane and a volumetric grid. This combination facilitates efficient local feature aggregation and feature alignment from both partial point clouds and multi-view images.

4.1. Triplane Variational Auto-Encoder

To conduct a latent triplane diffusion, the first step is to learn an encoder capable of encoding shapes in a triplane latent space. The Triplane VAE [19] comprises an encoder ψ_{enc} and a decoder ψ_{dec}. The encoder processes inputs and encodes them into latent space, while the decoder recovers shape from the latent representation. Surface point cloud $\mathbf{P} \in \mathbb{R}^{K\times 3}$ sampled from a ground truth shape, serves as inputs to ψ_{enc}. $K = 20,000$ in our experiment. These points are projected onto a triplane and subsequently processed by

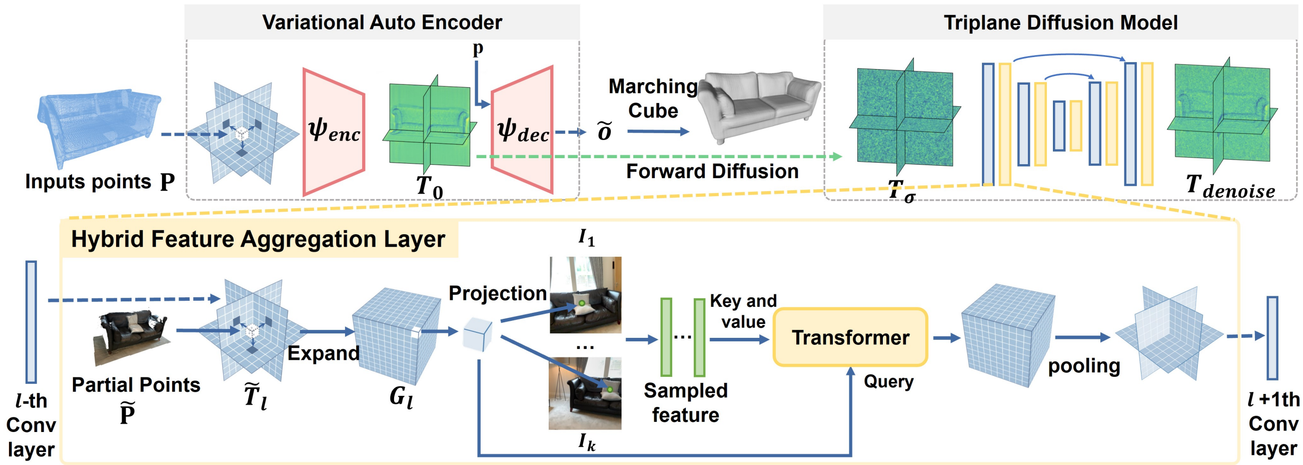

Figure 3. Pipeline of our DisCo. Firstly, a triplane VAE model is trained to encode shape into triplane latent space (top-left). Subsequently, a triplane diffusion model is trained in this latent space for conditional shape reconstruction (top-right). A novel Hybrid Feature Aggregation Layer is proposed to effectively aggregate and align local features in both partial points cloud and multi-view images (bottom).

a PointNet to form a triplane feature map. Followed by a UNet [41], the encoder ψ_{enc} outputs a normal distribution $\mathcal{N}(\mu, \sigma^2)$, where $\mu, \sigma \in \mathbb{R}^{H\times W\times 3\times C}$. A latent triplane $T_0 \in \mathbb{R}^{H\times W\times 3\times C}$ can be sampled from this distribution. The above process can be summarized as:

$$T_0 \sim \mathcal{N}(\mu, \sigma^2) \qquad \mu, \sigma = \psi_{enc}(\mathbf{P}) \tag{1}$$

The occupancy field is chosen as the shape representation. The decoder ψ_{dec} takes a point p in space and the triplane latent T_0 as inputs, yielding the occupancy of this point. Specifically, A UNet model first refines T_0 and outputs a triplane feature map. Then, point p is projected onto this triplane feature map, where features are sampled using bilinear interpolation. These sampled features are fed into MLPs which outputs the occupancy prediction $\tilde{o}$. The training of the VAE model is supervised by reconstruction loss and KL divergence loss as $\mathcal{L}_{vae} = \|\tilde{o} - o_{gt}\|_2^2 + \lambda_{kl}\mathcal{L}_{kl}$. In our experiment, $\lambda_{kl} = 0.025$. As in [51], the network layers in both ψ_{enc} and ψ_{dec} adopt 3d aware convolution for triplane processing.

The reconstruction is conducted in canonical object space. During training, augmentation such as random shifting, rotating and scaling are applied, so that it will be more robust to inaccurate objects' pose during inference.

4.2. Triplane Diffusion for reconstruction

Our approach employs latent triplane diffusion to reconstruct shapes based on partial point clouds and multi-view images as conditions. Specifically, a 2D UNet model serves as a denoise function $D(\cdot)$. This model comprises cascades of residual convolutional block and Hybrid Feature Aggregation Layer with 3d aware convolution. we introduce a novel Hybrid Feature Aggregation Layer, designed to foster effective interaction between local features derived from partial point clouds and multi-view images. We use continuous diffusion steps as in [24]. The forward diffusion process during training is defined as $T_\sigma = T_0 + n$, where $n \sim \mathcal{N}(0, \sigma^2\mathbf{I})$, σ indicates the diffusion steps, and also the deviation of the noise added. The denoise UNet takes a noisy triplane latent T_σ as input, and outputs a denoised triplane latent $T_{denoise}$. We follow the diffusion formulation in EDM [24], the objective function during training is:

$$\mathbb{E}_{\sigma,n,T_0}\lambda_\sigma\|D(T_0 + n, \sigma, c) - T_0\|_2^2 \quad n \sim \mathcal{N}(0, \sigma^2\mathbf{I}) \tag{2}$$

Here c denotes the conditional inputs. In our case, they are partial point clouds and multi-view images. λ_σ denotes a loss normalization factor. Inference can be conducted through sampling from noise triplane as in EDM [24].

During training our diffusion model, we implement a strategy to first pretrain on synthetic datasets, then fine-tune on LASA dataset. We leverage ShapeNet [5], ABO [10], and 3D-Future [15] dataset to synthesize partial point cloud and render images by embedding CAD models in HM3D scene[38, 60]. This strategy enhances both the robustness and the performance of our method, enabling more effective real-world object reconstruction.

4.3. Hybrid Feature Aggregation Layer

The objective of the Hybrid Feature Aggregation Layer is to aggregate and align local features from two input modalities to the triplane space. Points-triplane interaction is implemented by projecting points onto the triplane. However, such interaction becomes challenging when fusing images and triplanes. An alternative way is to expand the triplane into a volumetric grid and project the grids to images. This motivates us to propose this Hybrid Feature Aggregation

Table 2. Quantitative comparison on shape reconstruction. Evaluation metrics are mIoU / Chamfer L2 / F-score respectively. Higher mIoU and F-score are better while lower Chamfer L2 is better. Chamfer L2 is scaled by 1,000. LAS doesn't have mIoU value since it uses a surface occupancy representation.

Method	Chair	Sofa	Table	Cabinet	Bed	Shelf
IFNet	28.9 / 19.8 / 25.2	66.1 / 3.61 / 31.1	28.3 / 21.4 / 27.4	73.8 / 3.20 / 36.6	**64.0** / 2.95 / 28.8	17.2 / 5.24 / 29.6
LAS-pts	- / 13.2 / 22.6	- / 4.47 / 22.9	- / 19.5 / 23.9	- / 4.81 / 22.1	- / 4.77 / 25.8	- / 4.70 / 26.8
LAS-pts+imgs	- / 10.7 / 24.1	- / 3.76 / 24.1	- / 15.7 / 24.6	- / 4.56 / 24.4	- / 4.58 / 26.5	- / 4.35 / 27.2
3DShape2VecSec	32.5 / 6.43 / 25.6	66.7 / 3.39 / 27.6	30.9 / 12.9 / 25.3	68.1 / 4.09 / 23.7	62.3 / 3.69 / 30.5	22.3 / 3.79 / 32.0
Ours-pts	33.4 / 5.68 / 26.6	68.8 / 3.21 / 29.0	38.0 / 10.5 / 32.2	72.3 / 3.63 / 36.5	54.0 / 2.97 / 34.1	23.2 / 3.68 / 36.5
Ours-pts+imgs	**38.6 / 3.57/ 31.0**	**70.7/ 2.88/ 31.6**	**41.5 / 6.52 / 36.1**	**75.1 / 3.10 / 37.0**	62.5 / **2.62 / 35.4**	**24.5 / 3.45 / 37.5**

Layer, introducing a hybrid representation that facilitates both points-triplane and images-triplane interactions.

As in the bottom part of Fig. 3, inputs of this layer are partial point cloud $\tilde{P}$, k posed images $I_i, i \in 1, 2...k$, and triplane feature $\tilde{T}_l$ from the l-th convolutional layer. It begins with the aggregation of local features from partial points, achieved by projecting them onto the triplane and incorporating a PointNet layer. Image features map is first extracted using pretrained Vision Transformer. Then, the triplane expands into a volumetric grids $G_l \in \mathbb{R}^{H\times\ W\times L\times C}$. The volumetric grids are then projected to k images using their camera poses, and image local features are sampled using bi-linear interpolation.

To fuse features from multiple images, we employ a transformer. Specifically, the voxel feature serves as the query, while the sampled image feature serves as the key and value. The transformer outputs a volumetric grid attended with local image features from multi-view images. Finally, the volumetric grid is flattened back to a triplane by pooling, producing an output triplane feature map. This process ensures the effective aggregation and alignment of local features from diverse input modalities.

5. Occupancy-guided 3D Object Detection

3D object detection takes scene scans as input and parses the scene objects into 3D bounding boxes. In real-world scenarios, object scans are often incomplete and sparse due to occlusion, inaccurate sensors, and limited views during capture, making objects hard to recognize. To address it, we propose an Occupancy-Guided 3D Object Detection (OccGOD) approach that utilizes shape completeness prior for better scene understanding. We generate scene-level occupancy ground truth from LASA's fully-covered annotations. Specifically, scenes are partitioned into numerous 384x384x96 voxel grids, with a resolution of 4cm. All CAD models are then placed back into the scene, on which the surface points are densely sampled at 2cm intervals. Subsequently, we iterate through each point, marking their corresponding voxel to be occupied.

Our methodology follows a simple 'plug-and-play' manner. It is compatible with any detection method based on a 3D structured representation like volumetric grids [42, 49]. For ease of use, we built our OccGOD upon Cagroup3D [49]. In addition to the original backbone and bounding box prediction head, we introduce an occupancy head and augment the bounding box prediction head with another two output parameters for orientation regression. The occupancy heads takes backbone features as inputs, and output the occupancy of each voxel. More details of network design are in the supplemental. To further explore occupancy representations, we concatenate the features from the occupancy head with the backbone's features during ROI pooling for second-stage bounding box prediction. The occupancy head is supervised with a binary cross-entropy loss function with scene-level occupancy labels from LASA.

Our proposed OccGOD predicts a complete foreground structure to guide the bounding box detection. By leveraging the complete scene context, it achieves significant gains in detecting occluded and sparse objects compared to baseline methods.

6. Experiment

6.1. Experiment Setup

For indoor object reconstruction, we employ mean Intersection over Union (mIoU), L2 chamfer distance, and 1% F-score. We first normalize both results and ground-truth CAD into the interval from -0.5 to 0.5, and compute the above metrics between them. The experiment is conducted over 6 categories merged from 17 categories in LASA dataset.

For 3D object detection, we assess mean average precision (mAP) and mean average recall (mAR) with an IoU threshold at 0.5. Evaluations for both object reconstruction and 3D object detection are performed on LASA's test set.

6.2. Evaluation on Indoor Object Reconstruction

In this session, we compare our methods with existing baselines. We choose IFNet [8] and two state-of-the-art diffusion-based methods, LAS Diffusion [64] and 3DShape2VecSet [61], for comparison. We first pretrain all methods on synthetic dataset [5, 10, 15], then finetune them

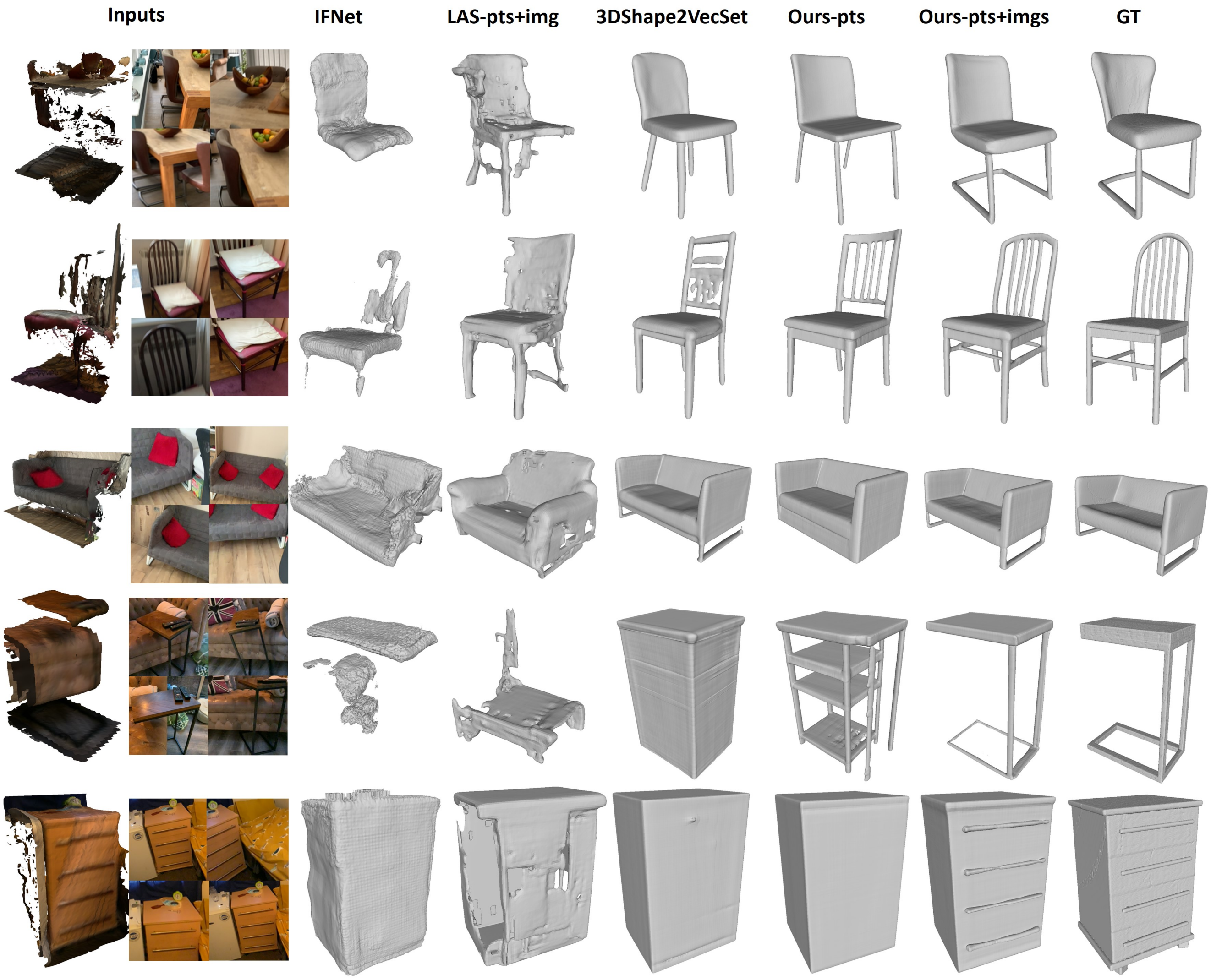

Figure 4. Qualitative comparison between our method and IFNet, 3DShape2Vecset, and LAS.

on LASA. IFNet and 3DShape2VecSet receive point clouds as inputs. For LAS Diffusion, we extend it into two versions: with point clouds as input (LAS-pts), and with both point clouds and images as input (LAS-pts+imgs). We compare them with our DisCo with two variants: 1) with partial point clouds only (Ours-pts); 2) with both modalities (Ours-pts+imgs). The quantitative and qualitative comparisons are shown in Tab. 2 and Fig. 4. Our method achieves state-of-the-art performance both quantitatively and qualitatively.

6.3. Ablation Study

Real-world Performance Boost using LASA We investigate the impact of our LASA dataset on real-world object reconstruction through three training setups: training from scratch on LASA (w/o pretrain), training on synthetic datasets only (w/o finetune), and pretraining on synthetic datasets followed by finetuning on LASA (full). The quantitative comparison is in Tab. 4. The table shows that finetuning on LASA significantly improves real-world reconstruction, with pretraining on synthetic data followed by finetuning on real strategy achieving the best performance.

Effectiveness of Hybrid Feature Aggregation Layer We verify the effectiveness of the Hybrid Feature Aggregation (HFA) layer in aggregating and aligning local features from both partial points and multi-view images. We compare it against not using the expanded grids to project onto the images. Specifically, the latter directly projects the triplane's pixels to the images. The quantitative comparison is shown in Tab. 5. The decent improvement of the HFA layer verifies its powerfulness in fusing multi-modal features.

Robustness to inaccurate detection We further investigate how inaccurate detection results could affect the re-

Table 3. Quantitative comparison between the state-of-the-art (CAGroup3D) and our OccGOD. Oriented bounding boxes are predicted. Evaluation metrics are mAP / mAR. Higher mAP / mAR indicates better performance.

$@IoU > 0.50$	Chair	Table	Cabinet	Refrigerator	Shelf	Bed	Sink	Washer	Bathtub
CaGroup3D	92.05 / 93.09	46.68 / 66.37	33.93 / 54.64	**90.56** / 91.22	35.12 / 57.54	67.31 / 73.89	64.03 / 74.15	**87.77 / 89.39**	**27.99** / 45.97
Our OccGOD	**92.57 / 93.25**	**49.17 / 66.76**	**35.05 / 55.90**	90.45 / **91.89**	**38.15 / 58.19**	**70.78 / 78.33**	**70.30 / 78.74**	86.27 / 88.64	26.56 / **46.77**
$@IoU > 0.50$	Toilet	Oven	Dishwasher	Fireplace	Stool	TV Monitor	Sofa	Stove	Overall
CAGroup3D	50.94 / 68.99	78.93 / 81.28	**92.19 / 92.86**	28.21 / 45.19	70.69 / 80.78	**1.16 / 6.63**	46.32 / **65.51**	33.32 / 41.73	55.72 / 66.43
Our OccGOD	**59.41 / 74.68**	**79.54 / 82.19**	90.96 / 92.75	**30.22 / 46.15**	**71.61 / 80.78**	0.73 / 5.19	**47.65** / 64.98	**35.64 / 41.73**	**57.36 / 67.47**

Table 4. Quantitative comparison on different training setups. The evaluation metrics are mIoU / chamfer L2 / F-score respectively.

Strategy	Chair	Sofa	Table
w/o pretrain	35.6 / 4.35 / 27.4	66.7 / 3.80 / 26.0	36.7 / 8.20 / 30.4
w/o finetune	25.0 / 10.8 / 23.5	66.7 / 5.11 / 26.8	30.9 / 12.5 / 26.1
full	**38.6 / 3.57/ 31.0**	**70.7/ 2.88/ 31.6**	**41.5 / 6.52 / 36.1**
	Cabinet	Bed	Shelf
w/o pretrain	72.6 / 3.49 / 34.2	**66.0 / 2.21** / 30.6	14.8 / 4.76 / 30.5
w/o finetune	67.2 / 5.22 / 29.1	55.3 / 3.38 / 29.0	23.9 / 4.18 / 30.8
full	**75.1 / 3.10 / 37.0**	62.5 / 2.62 / **35.4**	**24.5 / 3.45 / 37.5**

Table 5. Quantitative comparison between HFA layer and direct triplane projection. The evaluation metrics are mIoU / chamfer L2 / F-score respectively.

Method	Chair	Sofa
Triplane project	38.1 / 3.95 / 30.6	70.0 / 3.03 / 30.8
HFA layer	**38.6 / 3.57 / 31.0**	**70.7 / 2.88 / 31.6**

Table 6. Quantitative comparison between reconstruction using GT 3D bounding boxes and noisy bounding boxes. Evaluation metrics are mIoU / chamfer L2 / F-score respectively.

Detection type	Chair	Sofa	Table
noisy bbox	38.7 / 3.77 / 31.3	70.2 / 2.90 / 31.8	40.6 / 7.42 / 34.4
GT bbox	38.6 / 3.57 / 31.0	70.7/ 2.88/ 31.6	41.5 / 6.52 / 36.1
	Cabinet	Bed	Shelf
noisy bbox	73.1 / 3.70 / 34.4	65.5 / 2.64 / 36.2	24.1 / 3.85 / 36.0
GT bbox	75.1 / 3.10 / 37.0	62.5 / 2.62 / 35.4	24.5 / 3.45 / 37.5

construction. An experiment is conducted by randomly rotating between -10 and 10 degrees, scaling between 0.8 and 1.1, and shifting the center between -10% and 10% for each object. The quantitative results are shown in Tab. 6. We observe that, with considerable disturbances on object poses, our method achieves robust accuracy.

Effectiveness of scene-level occupancy to OccGOD Tab. 3 compares the baseline (Cagrounp3D) and our proposed OccGOD. Cagrounp3D achieves an mAP of 55.72 and an mAR of 66.43 with an IOU threshold of 0.5. Our OccGOD enhances the baseline, improving mAP by 1.64 and mAR by 1.04. Notable increases occurred for larger furniture like tables (+2.42 in AP), toilets (+8.47 in AP and +5.69 in AR), shelves (+3.03 in AP), beds (+3.47 in AP and 4.44 in AR), and sinks (+6.27 in AP and 4.59 in AR).

7. Conclusion

We have introduced a new dataset LASA, a Large-scale Aligned Shape Annotation Dataset. In this work, we have illustrated the substantial benefits LASA brings to the community, particularly in the realms of indoor instance-level scene reconstruction and 3D object detection. Empowered by LASA, we propose a novel Diffusion-based Cross-Modal Shape Reconstruction approach, namely DisCo, and an Occupancy-guided 3D Object Detection method, namely OccGOD. In DisCo, we design a novel Hybrid Feature Aggregation Layer to effectively fuse and align local features from two input modalities - partial point clouds and multi-view images. In OccGOD, we leverage the scene-level occupancy labels provided by LASA, to enhance 3D object detection by learning object completeness priors. Extensive experiments demonstrated that, with the support of LASA, both methods achieve state-of-the-art performance in real-world scenarios.

We firmly believe that the large-scale and well-aligned features of LASA present better annotation quality and quantity, laying a foundation for many 3D downstream applications, including 3D understanding and reconstruction.

8. Acknowledgements

The work was supported in part by Guangdong Provincial Outstanding Youth Project No. 2023B1515020055, the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, NSFC-62172348, the National Key R&D Program of China with grant No. 2018YFB1800800, by Shenzhen Outstanding Talents Training Fund 202002, by Guangdong Research Projects No. 2017ZT07X152 and No. 2019CX01X104, by Key Area R&D Program of Guangdong Province (Grant No. 2018B030338001), by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), and by Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. ZDSYS201707251409055). It is also partly supported by NSFC-61931024 and Shenzhen General Project No. JCYJ20220530143604010.

References

[1] Himanshu Arora, Saurabh Mishra, Shichong Peng, Ke Li, and Ali Mahdavi-Amiri. Multimodal shape completion via implicit maximum likelihood estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2958–2967, 2022. 2, 3

[2] 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/CVF Conference on computer vision and pattern recognition*, pages 2614–2623, 2019. 2, 3, 4

[3] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 2

[4] Miguel Angel Bautista, Walter Talbott, Shuangfei Zhai, Nitish Srivastava, and Joshua M Susskind. On the generalization of learning-based 3d reconstruction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2180–2189, 2021. 3

[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. 2, 3, 5, 6

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

[7] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4456–4465, 2023. 2, 3, 4

[8] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6970–6981, 2020. 1, 3, 6

[9] 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 *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 628–644. Springer, 2016. 2, 3

[10] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21126–21136, 2022. 2, 5, 6

[11] 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. 2

[12] 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. 3

[13] Mingyue Dong, Linxi Huan, Hanjiang Xiong, Shuhan Shen, and Xianwei Zheng. Shape anchor guided holistic indoor scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21916–21926, 2023. 2, 3

[14] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021. 2

[15] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129:3313–3337, 2021. 2, 5, 6

[16] Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 484–499. Springer, 2016. 3, 4

[17] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–224, 2018. 3

[18] Can Gümeli, Angela Dai, and Matthias Nießner. Roca: Robust cad model retrieval and alignment from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4022–4031, 2022. 2, 3

[19] Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, and Barlas Oğuz. 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv preprint arXiv:2303.05371*, 2023. 2, 3, 4

[20] Christian Häne, Shubham Tulsiani, and Jitendra Malik. Hierarchical surface prediction for 3d object reconstruction. In *2017 International Conference on 3D Vision (3DV)*, pages 412–420. IEEE, 2017. 3

[21] 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. 3, 4

[22] Muhammad Zubair Irshad, Thomas Kollar, Michael Laskey, Kevin Stone, and Zsolt Kira. Centersnap: Single-shot multi-object 3d shape reconstruction and categorical 6d pose and size estimation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 10632–10640. IEEE, 2022.

[23] Muhammad Zubair Irshad, Sergey Zakharov, Rares Ambrus, Thomas Kollar, Zsolt Kira, and Adrien Gaidon. Shapo: Im-

plicit representations for multi-object shape appearance and pose optimization. 2022. 3

[24] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022. 5

[25] Andrey Kurenkov, Jingwei Ji, Animesh Garg, Viraj Mehta, JunYoung Gwak, Christopher Choy, and Silvio Savarese. Deformnet: Free-form deformation network for 3d shape reconstruction from a single image. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 858–866. IEEE, 2018. 3

[26] Or Litany, Alex Bronstein, Michael Bronstein, and Ameesh Makadia. Deformable shape completion with graph convolutional autoencoders. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1886–1895, 2018. 3

[27] Haolin Liu, Yujian Zheng, Guanying Chen, Shuguang Cui, and Xiaoguang Han. Towards high-fidelity single-view holistic reconstruction of indoor scenes. In *European Conference on Computer Vision*, pages 429–446. Springer, 2022. 2, 3

[28] Kevis-Kokitsi Maninis, Stefan Popov, Matthias Nießner, and Vittorio Ferrari. Vid2cad: Cad model alignment using multi-view constraints from videos. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):1320–1327, 2022. 2, 3

[29] Kevis-Kokitsi Maninis, Stefan Popov, Matthias Nießner, and Vittorio Ferrari. Cad-estate: Large-scale cad model annotation in rgb videos. *arXiv preprint arXiv:2306.09011*, 2023. 2, 3, 4

[30] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 3

[31] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 55–64, 2020. 3

[32] 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, 3

[33] Junyi Pan, Xiaoguang Han, Weikai Chen, Jiapeng Tang, and Kui Jia. Deep mesh reconstruction from single rgb images via topology modification networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9964–9973, 2019. 3

[34] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20133–20143, 2023. 2, 3, 4

[35] 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/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 3

[36] 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. 1

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

[38] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 5

[39] Siddhant Ranade, Christoph Lassner, Kai Li, Christian Haene, Shen-Chi Chen, Jean-Charles Bazin, and Sofien Bouaziz. Ssdnerf: Semantic soft decomposition of neural radiance fields. *arXiv preprint arXiv:2212.03406*, 2022. 2, 3, 4

[40] Jason Rock, Tanmay Gupta, Justin Thorsen, JunYoung Gwak, Daeyun Shin, and Derek Hoiem. Completing 3d object shape from one depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2484–2493, 2015. 3

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 5

[42] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Fcaf3d: Fully convolutional anchor-free 3d object detection. In *European Conference on Computer Vision*, pages 477–493. Springer, 2022. 2, 6

[43] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019. 3

[44] J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20875–20886, 2023. 2, 3

[45] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2

[46] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2974–2983, 2018. 2, 3, 4

[47] Jiaxiang Tang, Xiaokang Chen, Jingbo Wang, and Gang Zeng. Point scene understanding via disentangled instance mesh reconstruction. In *European Conference on Computer Vision*, pages 684–701. Springer, 2022. 2, 3

[48] Michał J Tyszkiewicz, Kevis-Kokitsi Maninis, Stefan Popov, and Vittorio Ferrari. Raytran: 3d pose estimation and shape reconstruction of multiple objects from videos with ray-traced transformers. In *European Conference on Computer Vision*, pages 211–228. Springer, 2022. 2, 3

[49] Haiyang Wang, Shaocong Dong, Shaoshuai Shi, Aoxue Li, Jianan Li, Zhenguo Li, Liwei Wang, et al. Cagroup3d: Class-aware grouping for 3d object detection on point clouds. *Advances in Neural Information Processing Systems*, 35: 29975–29988, 2022. 2, 6

[50] 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. 2, 3

[51] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4563–4573, 2023. 3, 4, 5

[52] Rundi Wu, Xuelin Chen, Yixin Zhuang, and Baoquan Chen. Multimodal shape completion via conditional generative adversarial networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 281–296. Springer, 2020. 2, 3

[53] Yushuang Wu, Zizheng Yan, Ce Chen, Lai Wei, Xiao Li, Guanbin Li, Yihao Li, Shuguang Cui, and Xiaoguang Han. Scoda: Domain adaptive shape completion for real scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17630–17641, 2023. 2, 3, 4

[54] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE winter conference on applications of computer vision*, pages 75–82. IEEE, 2014. 2, 3

[55] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2690–2698, 2019. 2, 3

[56] Haozhe Xie, Hongxun Yao, Shengping Zhang, Shangchen Zhou, and Wenxiu Sun. Pix2vox++: Multi-scale context-aware 3d object reconstruction from single and multiple images. *International Journal of Computer Vision*, 128(12): 2919–2935, 2020. 1, 3

[57] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *Advances in neural information processing systems*, 32, 2019. 3

[58] Xingguang Yan, Liqiang Lin, Niloy J Mitra, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Shapeformer: Transformer-based shape completion via sparse representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6239–6249, 2022. 1, 2, 3

[59] Zhenpei Yang, Zhile Ren, Miguel Angel Bautista, Zaiwei Zhang, Qi Shan, and Qixing Huang. Fvor: Robust joint shape and pose optimization for few-view object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2497–2507, 2022. 3

[60] Zhenpei Yang, Zaiwei Zhang, and Qixing Huang. Hm3d-abo: A photo-realistic dataset for object-centric multi-view 3d reconstruction. *arXiv preprint arXiv:2206.12356*, 2022. 5

[61] Biao Zhang, Jiapeng Tang, Matthias Nießner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Trans. Graph.*, 42(4), 2023. 1, 2, 3, 4, 6

[62] Cheng Zhang, Zhaopeng Cui, Yinda Zhang, Bing Zeng, Marc Pollefeys, and Shuaicheng Liu. Holistic 3d scene understanding from a single image with implicit representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8833–8842, 2021. 3

[63] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Proceedings of The European Conference on Computer Vision (ECCV)*, 2020. 2

[64] Xin-Yang Zheng, Hao Pan, Peng-Shuai Wang, Xin Tong, Yang Liu, and Heung-Yeung Shum. Locally attentional sdf diffusion for controllable 3d shape generation. *ACM Transactions on Graphics (SIGGRAPH)*, 42(4), 2023. 2, 3, 4, 6