Gen3DSR: Generalizable 3D Scene Reconstruction via Divide and Conquer from a Single View

Andreea Ardelean Mert Özer Bernhard Egger
Friedrich-Alexander-Universität Erlangen-Nürnberg
{andreea.dogaru, mert.oezer, bernhard.egger}@fau.de

Figure 1. Our method can reconstruct a 3D scene from a single view. We identify distinct objects, address their occlusions through amodal completion, and reconstruct them individually. The resulting 3D objects are composed into the scene using monocular depth guides. Each component is reconstructed as a triangle mesh, enabling downstream applications such as scene manipulation and editing.

Abstract

Single-view 3D reconstruction is currently approached from two dominant perspectives: reconstruction of scenes with limited diversity using 3D data supervision or reconstruction of diverse singular objects using large image priors. However, real-world scenarios are far more complex and exceed the capabilities of these methods. We therefore propose a hybrid method following a divide-and-conquer strategy. We first process the scene holistically, extracting depth and semantic information, and then leverage an object-level method for the detailed reconstruction of individual components. By splitting the problem into simpler tasks, our system is able to generalize to various types of scenes without retraining or fine-tuning. We purposely design our pipeline to be highly modular with independent, self-contained modules, to avoid the need for end-to-end training of the whole system. This enables the pipeline to naturally improve as future methods can replace the individual modules. We demonstrate the reconstruction performance of our approach on both synthetic and real-world scenes, comparing favorable against prior works. Project page: https://andreeadogaru.github.io/Gen3DSR

1. Introduction

Single-view 3D scene reconstruction refers to the problem of understanding and explaining all the visible components that assembled together create a 3D scene which closely reproduces the original 2D observation. The computer vision and graphics communities have long been interested in automating this task, yet its complexity still leaves room for many improvements [37, 71]. Successful single-view applications have been developed for specific purposes such as face reconstruction [17, 36] and hair modeling [75]. However, 3D understanding from a single image is far from solved in the case of larger scale problems such as indoor/outdoor scene reconstruction with multiple objects [65].

In general, even reconstructing one 3D object from a single image is a severely ill-posed problem, *e.g.* it is impossible to tell precisely how the back side of an object looks like if the input image only observes the front. Nonetheless, if the distribution of objects that are naturally present in our day-to-day lives is known, one can plausibly predict the shape and appearance of a 3D object from very limited information. Accordingly, various priors have been used in the context of particular object classes (such as simple shapes [2], or

human faces [17]). However, modeling entire scenes is a significantly more challenging problem.

Given the complexity of real-world scenes, reversing the process of image capturing in an end-to-end fashion would require a huge amount of data covering the variability of realistic environments. Therefore, many works solve a simplified version of the task by focusing on single objects or indoor rooms with a limited number of object classes. Under these assumptions, most of the existent solutions rely on 3D scene geometry supervision from synthetic datasets. This class of methods usually struggles when applied to real-world images due to the domain gap and limited diversity in existing datasets. In contrast, we propose to tackle the single-view 3D scene reconstruction problem in a divide-and-conquer approach while building on the advances in related, simpler tasks. In Figure 1 we show that, following this approach, our pipeline is able to reconstruct multi-object scenes with unprecedented quality.

In the past few years, the field of computer vision has seen tremendous progress in solving particular tasks such as depth estimation and single-image 3D object reconstruction. It is the right time for these components to be assembled to solve the challenging task of full 3D scene reconstruction. We have identified the following sub-problems that together comprehensively explain a 3D scene and enable its reconstruction from a single input image: estimating the camera calibration, predicting the (metric) depth map, segmenting entities, detecting foreground instances, reconstructing the background, recovering the occluded parts of the individual objects (amodal completion), and reconstructing them. Our disentangled framework is open for incremental improvements and future enhanced modules can be easily plugged in to boost the reconstruction performance of the entire system. The main contributions of this paper are as follows:

- We design a compositional framework and the corresponding abstractions, enabling scene-level 3D reconstruction without end-to-end training.
- We build a model for amodal completion and show how it can be used towards achieving full scene reconstruction.
- We develop the connecting links for integrating individually reconstructed 3D objects into the scene layout by exploiting single-view depth estimation.
- We achieve an unmatched level of generalizability for real-world single-view 3D scene reconstruction, which we demonstrate through extensive evaluations.

2. Related Work

Starting with the pioneering work of Lawrence Roberts [62], the task of recovering 3D scene properties such as geometry, texture, and layout from a single 2D image has been extensively studied. Learning-based advancements led to substantial progress in this challenging task. However, existing methods still have limited understanding and struggle to faithfully reconstruct complex realistic scenes. In this section, we review the major lines of work and highlight the advantages of our approach that enable state-of-the-art results for 3D scene reconstruction from a single view.

Feed-forward scene reconstruction. The direct regression conditioned on the input image used in many computer vision tasks can also be employed for 3D scene reconstruction. Provided the availability of large scene collections, such systems are trained end-to-end using 3D supervision. Most such works [11, 12, 85] rely on an encoder-decoder architecture that takes as input the image and predicts voxel grids containing scene properties such as geometry, semantics and instance labels. These methods have the advantage that by predicting the scene layout jointly with the containing objects, the object poses are intrinsically correct. However, these solutions suffer from resolution limitations, require large 3D data collections for training [19, 20] and usually do not generalize well to real-world scenes. In contrast, we use a modular framework that does not require end-to-end training or 3D supervision.

Factorized scene reconstruction. Following the formulation of Tulsiani *et al*. [72], the scene can be considered as composed of different factors, including layout, object shape and pose, that together make up a 3D scene representation. As this approach reconstructs the scene components individually, it is beneficial for downstream applications and the reconstruction system can be designed with specialized components for each type of factor. Methods following this paradigm initially reconstructed the objects in the scene as bounding boxes [16, 27, 67] or voxels [40, 43, 72], while later works use meshes [22, 28, 41, 55] or neural fields [45, 68, 83, 84]. Apart from the elementary factors of [72], more recent methods also include camera attributes [27, 28, 45, 55], textures [10, 79, 81], lightning [79, 81], or even material properties [79, 81]. Depending on the focus of the method, some works rely on a retrieval algorithm for identifying the object candidates [3, 28, 30, 41, 79] or regress their geometry [22, 45, 68, 84]. The initially predicted scene components are sometimes further fine-tuned using derivative-free optimization [9, 28, 30] or differentiable rendering [22, 79, 81, 83]. Our method also uses a compositional scene reconstruction framework; in our pipeline, each module is trained individually on specific datasets. This strategy allows our proposed solution to transcend predefined (limited) classes of objects and reconstruct diverse scenes effectively.

Single-view scene understanding. The reconstruction of simple scenes composed of a single isolated object has seen numerous approaches along the years [24]. Nonetheless, enabled by the versatile diffusion-based models capable of generating realistic images [57, 64] and large collections of 3D objects [13, 14], there has recently been a surge of single-view 3D reconstruction methods [46–48, 69] capa-

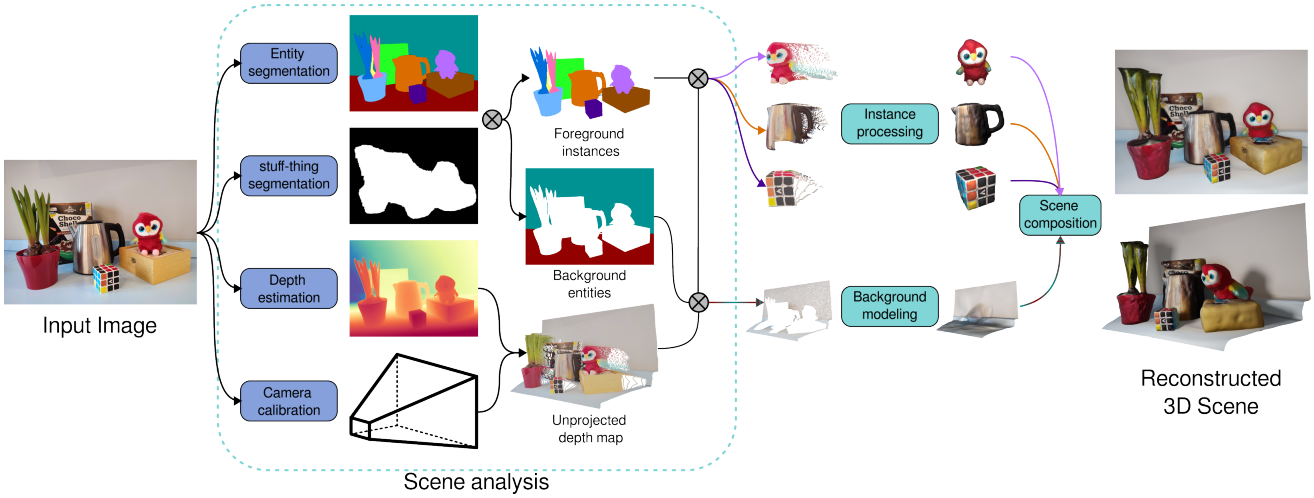


Figure 2. Method Overview: the input image is first analyzed collectively by an ensemble of state-of-the-art monocular models. Subsequently, the identified instances are individually processed, as elaborated in Figure 3. The reconstructed objects, along with the modeled background, are composed into the final scene, which can then be used in various applications.

ble of creating digital objects with unprecedented quality. Closely related to 3D scene reconstruction is monocular depth estimation [51], which predicts from a single image a 2.5D representation of the visible content. Recent methods focus on improving the generalizability of the estimator by following a training protocol employing large datasets [5, 80] or by fine-tuning a large image prior such as Stable Diffusion [64] for the depth estimation task [34]. A scene is also described by the camera used to capture the image [25, 33, 42] and by the separation between different instances [38, 60, 82]. Though the current state-of-the-art methods for each individual sub-task achieve robust results for real images, they only offer an incomplete explanation of the scene. Therefore, we compose them in an open framework capable of fully reconstructing complex real-world scenes from a single image.

3. Method

Our method, as illustrated in Figure 2, takes as input a single RGB image I and predicts the full 3D scene reconstruction $R(I)$ represented as a collection of triangle meshes. The proposed solution does not require end-to-end training and instead relies on off-the-shelf models carefully integrated into a seamless framework. First, we parse the image of the scene by finding the composing entities, and estimating the depth and camera parameters. Then, we separate the identified entities in *stuff* (amorphus shapes) and *things* (characteristic shapes) [7]. To recover the full view of each object, we perform amodal completion on the masked crops of the instances. Each object is reconstructed individually in a normalized space and aligned to the view space using the scene layout guides from the depth map. Importantly, we address the differences in focal length, principal point, and camera-to-object distance between the two spaces through reprojection. Finally, we model the background as the surface that approximates the stuff entities collectively.

3.1. Scene analysis

Our framework decouples the object-agnostic from the object-specific processing, pursuing a balance between the representational power and the generalization capabilities of the integrated modules: That is, the perspective properties, semantic labels, and depth information are best retrieved by perceiving the scene as a whole

The geometry of a scene is characterized by its layout and the amodal shape of the contained objects. The layout of a scene refers to the surfaces that enclose the space (*e.g.*, walls) and the 3D locations of the objects. To model the layout, we unproject a **monocular depth estimation** D, of the input image using predicted **camera calibration** parameters, K_{img}, as a point cloud in the 3D view space, $P^{view} \in \mathbb{R}^3$, and adopt it as our guide for positioning the scene components. A 2.5D representation is not sufficient to fully describe the layout of a scene as it only provides information for the visible parts. Still, it can be used to integrate individually reconstructed 3D objects into the scene (Section 3.2) and for background estimation (Section 3.3).

To parse the image, we opt for an **entity segmentation** [59, 60] approach in contrast to conventional 2D object detection, which enables us to segment all semantically-meaningful entities without being constrained to a predefined set of classes. This step partitions the image I into instances $\{M_i\}$ that can be individually reconstructed to compose the whole scene. Furthermore, we consider the natural separation of the instances in *thing* and *stuff* using a universal image segmentation model. This facilitates our method to tailor the reconstruction process for each group, leveraging their unique properties (objects vs background).

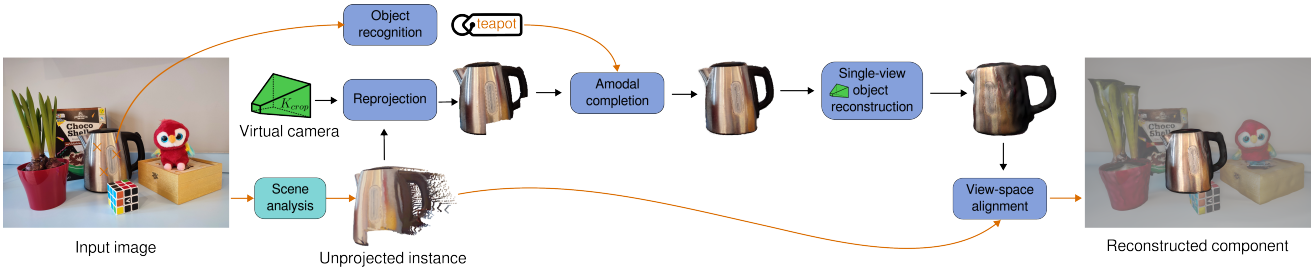


Figure 3. Detailed Overview over the processing steps of each instance. An image is processed through the scene analysis part of our framework as described in Figure 2. Then, we add an object recognition information for diffusion guided completion to restore partially occluded objects. Lastly, we perform reconstruction and align the result back to the input view space.

In this stage, we also label the identified entities, which can provide more context for instance processing.

3.2. Instance processing

Let an object O_i be a well-defined shape categorised as *thing*, identified by its entity mask M_i, with corresponding RGB-D region $I_i = I[M_i]$, $D_i = D[M_i]$, and a label L_i. The processing steps are illustrated in Figure 3. We reconstruct each instance individually to fully benefit from a view-conditioned 2D diffusion model $\mathcal{Z}$ [14, 48], trained using multi-view images of mostly single objects from large scale collections [13, 14]. As the images used to train these models are rendered with a fixed predefined camera configuration K_{crop}, they generalize poorly to in-the-wild object crops. Therefore, we propose to address the domain-shift via **reprojection** of the object-associated pixels. To this end, we identify the virtual camera that together with the desired intrinsics K_{crop} closely matches the observed image. Then, we use the transformation to project the unprojected pixels P_i^{view} to a crop C_i that resembles the training domain of $\mathcal{Z}$.

For simple scenes in which there is no occlusion between instances, the crop C_i represents the full view of the object O_i. However, this is not the case for most real-world scenes, and directly feeding C_i to the object reconstruction method $\mathcal{R}$ would result in an incomplete object. Therefore, we propose to recover the missing parts of C_i by leveraging the image prior embodied by pre-trained large-scale diffusion models like Stable Diffusion. We approach the task named **amodal completion** which deals with recovering the shape and appearance of partially visible instances as an image-to-image translation problem. Specifically, we train a model to predict the view $\hat{C}_i$ of the full object O_i conditioned on the object parts depicted in C_i and the label L_i. Given the difficulty of collecting well-segmented training images guaranteed to contain complete objects, we generate a synthetic dataset specifically for this task. We render synthetic objects from a large collection and then obtain the conditioning images by masking out parts of the object with a randomly overlaying silhouette of another object. For more details about the dataset generation and differences between amodal completion and inpainting please see the supplementary material. We use this dataset to fine-tune Stable Diffusion, following the methodology of InstructPix2Pix [6], and concatenate the encoded conditioning image to the noisy latent. The weights added to the base network are initialized with zero, while the rest are taken from the pre-trained model to benefit from the learned image prior.

The complete crop $\hat{C}_i$ can now be used as input for the **single-image 3D object reconstruction** method $\mathcal{R}$. Using a view-conditioned diffusion prior enables the model to reconstruct a wide range of objects from a single view without the need for 3D training supervision. The object is reconstructed using a differentiable 3D representation (*e.g.*, neural fields [52, 53] or 3D Gaussians [35]) that is either directly fitted to multi-view images generated by the diffusion prior, or by optimizing a Score Distillation Sampling-based loss [58] against the diffusion prior. Then, a polygonal mesh R_i^{obj} aligned with the input crop $\hat{C}_i$ is extracted from the 3D representation using Marching Cubes [49, 70]. The obtained mesh can optionally be further fine-tuned to refine the texture of the reconstructed object.

We transform the reconstructed instance from the object space (DreamGaussian) R_i^{obj} to the view space (our scene) with the inverse transformation determined by the virtual camera used for projection. The obtained mesh R_i^{view} is aligned to the object points P_i up to an unknown scale factor s_i; this is because $\mathcal{R}$ reconstructs objects at an arbitrary scale. We estimate s_i as the scale factor that minimizes the distance between the visible points in R_i^{view} and P_i^{view}.

3.3. Background modeling

Compared to the objects contained in a scene, the entities categorised as *stuff*, *e.g.* walls or ceiling, usually have a simpler geometry, which can be partially approximated by the corresponding regions in the depth map. Nonetheless, large areas are occluded by the foreground objects. We consider all the background instances as one described by their mask union M_{bg} and the corresponding image I_{bg} and depth data D_{bg}. Then, we fit one small Multi-layer Perceptron $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ to represent the signed distance function (SDF) to the unprojected background points P_{bg}^{view}, and another one $c : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ to model the color associated to them.

Lastly, we sample a dense grid of points in the camera frustum and evaluate their SDF values. The background surface is defined as $\mathcal{S} = \{x \in \mathbb{R}^3 | f(x) = 0\}$ and can be extracted from the grid using marching cubes [49, 70].

4. Experiments

4.1. Implementation details

As the proposed framework is not constrained to specific modules, we leverage the significant progress made by the computer vision community in the recent years towards solving the different sub-tasks described above.

In the scene analysis stage, we rely on CropFormer [59] for entity segmentation, OneFormer [31] to separate them in foreground instances and background entities, and Perspective Fields [33] to estimate the camera calibration. We mainly use Marigold [34] for depth estimation. However, the model predicts affine-invariant depth which differs by an unknown image-level offset and scale from the absolute physical units. During evaluations, we estimate these factors based on the ground truth depth available in the datasets to ensure that the reconstructions align with the target. For in-the-wild predictions, we empirically found that estimating the two unknowns of Marigold output based on a metric depth estimation, in our case, DepthAnything [80], achieves better results than using the latter by itself.

In the instance processing stage, we perform amodal completion on the reprojected crops using our model obtained by fine-tuning Stable Diffusion v1.5 [64]. We also sample several points in the instance's mask and feed them to OVSAM [82] together with the input image to obtain the text prompt for guiding the diffusion. The completed object is then reconstructed using DreamGaussian [69]. We estimate the camera elevation required by DreamGaussian as in [46]. Then, we find the 3D points of the reconstruction that correspond to the unprojected instance points, which serve as our layout guide, and compute the scale which aligns them. Further specifications regarding the integrated models, possible alternatives for some of the processing stages and an analysis of the inference time of our method are provided in Section 6 of the supplementary material.

4.2. Results

We showcase the performance of our method using several datasets across diverse scenarios. For numerical evaluation we first consider 3D-FRONT [19, 20], a synthetic dataset of indoor rooms with available ground truth geometry. Due to the large scale of the dataset, we manually sample 100 images from the test split of [45], avoiding the images with heavy scene occlusions (*e.g.*, camera positioned behind a plant), intersecting objects, and scenes with very few objects. In addition, we use the 10 validation images of HOPE-Image [73] dataset containing household objects captured under two scenarios. Though the dataset is intended for object pose evaluation, we find the ground-truth object alignment to match the input images well-enough for our purpose.

Method	3D-supervision-free	Zero-shot	Compositional
InstPIFu [45]	✗	✗	✓
USL [22]	✓	✗	✓
BUOL [11]	✗	✗	✗
Uni-3D [85]	✗	✗	✗
DreamG [69]	✓	✓	✗
Ours	✓	✓	✓

Table 1. As opposed to existing approaches, our method is at the same time compositional, able to generalize in a zero-shot manner, and does not require training with 3D supervision.

	3D-FRONT [19]		HOPE-Image [73]	
Method	Chamfer ↓	F-Score ↑	Chamfer ↓	F-Score ↑
InstPIFu [45]	0.124	**74.14**	-	-
DreamG [69]	0.207	41.09	3.038	29.57
DreamG + reprojection	0.187	49.98	4.059	30.62
Ours	**0.120**	68.82	**1.446**	**54.82**

Table 2. Quantitative results on foreground instances reconstruction. On 3D-FRONT, we are on par with InstPIFu, despite it being trained with 3D data on this dataset. On HOPE-Image, InstPIFu fails to recognize any of the evaluated objects. On both datasets, we outperform the non-compositional approach, DreamGaussian.

We report the quantitative results using the widely-employed metrics for 3D reconstruction: Chamfer Distance [4] and F-Score [39]. Both are computed between densely sampled sets of points from the reconstructed meshes and the ground truth respectively. As we focus the evaluation on whole scenes, the points are uniformly sampled from the entire geometry of a scene.

We compare our method against several solutions covering multiple approaches for 3D scene reconstruction. An overview of their capabilities is presented in Table 1. BUOL[11] and Uni-3D [85] are both feed-forward scene reconstruction methods that have been trained with 3D supervision on the 3D-FRONT dataset. While we evaluate the methods on the same dataset they were trained on, we use a more realistic rendering, following [45]. Even under this minor change, the methods' performance degrades significantly, as can be seen in Table 3 and Figure 4, showing their lack of generalization.

InstPIFu [45] and USL [22] are both compositional approaches that rely on 2D and 3D object detectors for identifying the objects to be reconstructed and aligning them in the scene. However, InstPIFu requires direct 3D supervision and USL is trained end-to-end. This limits their application domain and generalizability. Furthermore, as seen in Tables 3 and 2, our method is able to quantitatively match the performance of InstPIFu even when evaluated on the 3D-FRONT dataset (which InstPIFu used for training). The qualitative

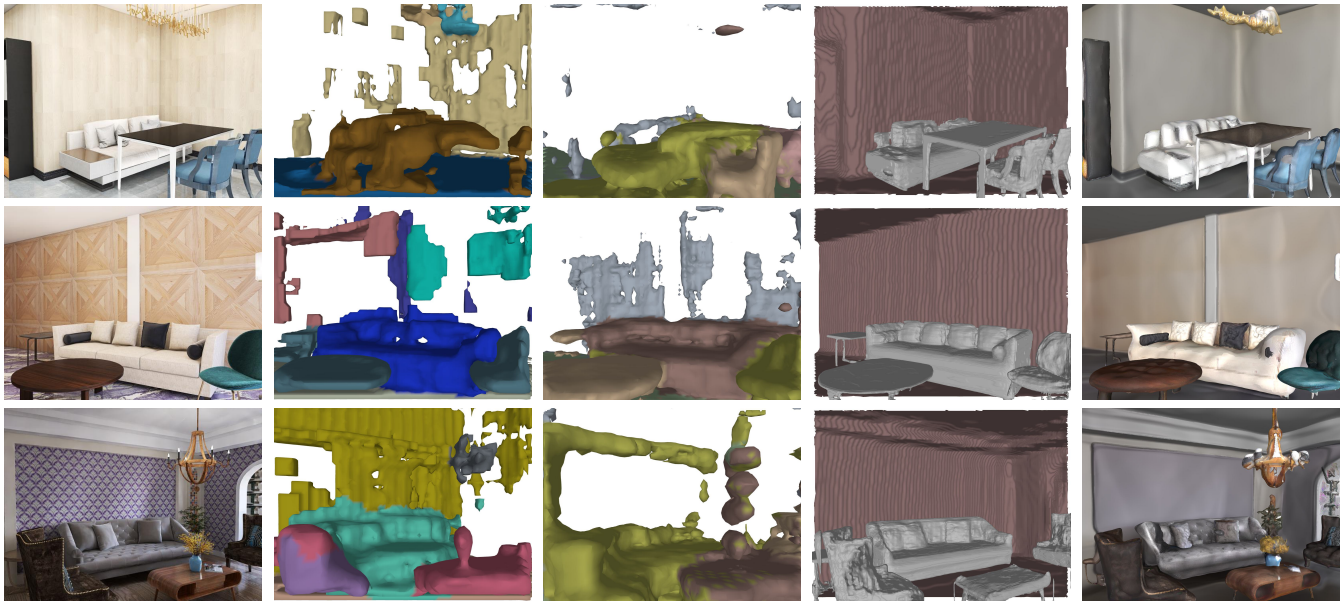


Figure 4. Qualitative results on the 3D-FRONT dataset [19]. The methods considered reconstruct full scenes including background regions and foreground instances.

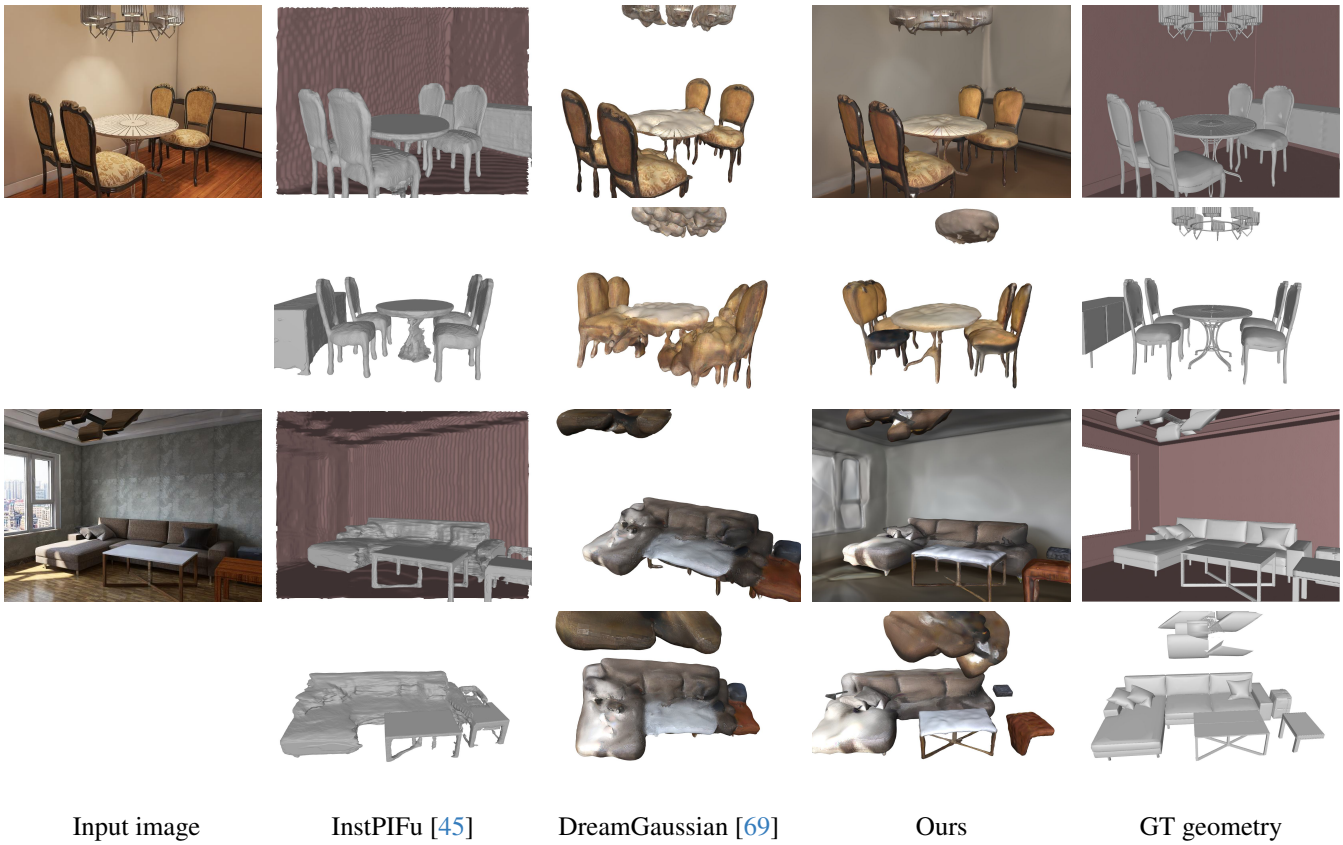


Figure 5. Qualitative results on 3D-FRONT [19] illustrated under the input view and a second one chosen to highlight the reconstruction performance. In contrast to Ours and InstPIFu, DreamGaussian reconstructs all the objects at once and does not model the background.

Method	Chamfer ↓	F-Score ↑
InstPIFu [45]	0.119	70.63
BUOL [11]	0.294	37.04
Uni-3D [85]	0.448	32.97
Ours	**0.099**	**75.33**

Table 3. Quantitative evaluation on full scene reconstruction including background and foreground elements on the 3D-FRONT dataset [19]. Our method, which performs in a zero-shot fashion, outperforms all the baselines, which have been trained using 3D supervision on a version of this dataset.

results in Figures 4 and 5 show that the method successfully reconstructs large furniture pieces and arranges them in a good layout; however, several objects such as plants and chandeliers are missing. The results of InstPIFu further degrade when evaluated out-of-distribution, as can be seen on real-world images in Figure 7. Since the implementation of USL [22] is not publicly available, we only compare our visual results on the Hypersim dataset [63] in Figure 6. USL also misses many objects, and the reconstructed components have a simplified geometry with no texture. In contrast, our results are more realistic and have higher visual quality.

Additionally, we compare our method with DreamGaussian [69] by itself in a non-compositional approach. To apply the model, we treat all instances in a scene as a single object and reconstruct them together. As the model has seen several scenes composed of more than one object during its training [14], it performs reasonably under this setting. Given the different camera intrinsics in the evaluation, we also compare the results of applying the model on the images after using a reprojection similar to the one used in our pipeline. This further boosts its performance as measured in Table 2. Still, its results are worse compared to our compositional approach, which can be analyzed in the qualitative results in Figure 5 and in the supplementary material.

Lastly, we evaluate in Figure 7 the methods' performance on real-world data with diverse scenarios. The results show that the proposed solution reconstructs complex scenes well, while overcoming many limitations of prior works. We briefly address the reconstruction of outdoor scenes in the Section 7.4 of the supplementary material.

4.3. Ablations

The two most important interfaces in combining the depth estimation and the single-view object reconstruction components are our reprojection and amodal completion. We present an ablation of these components in Table 4. Ablating the reprojection amounts to simply using image crops. This ignores projective geometry properties and leads to deformed reconstructions. Amodal completion is necessary to contend with the various occlusions that appear in the input view. This step boosts the performance on the 3D-FRONT dataset, but does not improve the numerical evaluation on

	3D-FRONT [19]		HOPE-Image [73]	
Method	Chamfer ↓	F-Score ↑	Chamfer ↓	F-Score ↑
Ours (full)	**0.120**	**68.82**	**1.446**	54.82
w/o reprojection	0.155	58.56	1.505	51.89
w/o amodal comp.	0.122	66.81	1.450	**54.99**

Table 4. Ablation results for our reprojection and amodal completion modules; both contribute to the performance of our method.

HOPE-Image dataset, since most of the objects in the scenes are not occluded. More ablation results are included in the Section 7.3 of the supplementary material.

4.4. Limitations

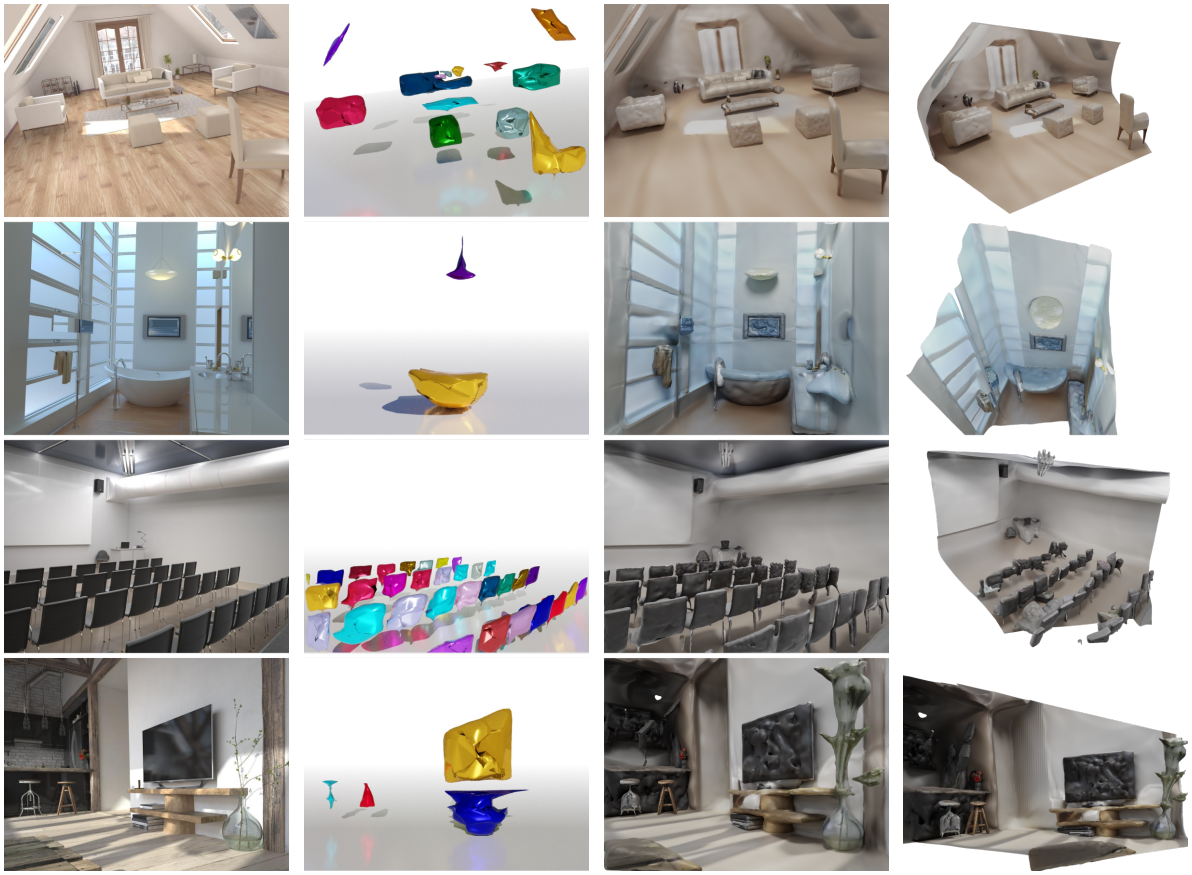
The proposed method has certain shortcomings and there is significant room for improvement, especially for in-the-wild predictions. By design, the failure cases of the individual modules (depth estimation, camera calibration, elevation estimation, etc.) become limitations of our framework. Since we do not train an end-to-end system, errors can propagate from one stage to the next. Therefore, the performance of the overall pipeline is limited by its weakest link.

We believe that most of the current limitations can be overcome by improving the implementation of some of the particular modules in our framework and by enhancing their interoperability: using estimated depth in 3D object reconstruction or global image context for amodal completion. We further discuss the method's limitations and provide concrete examples in the Section 8 of the supplementary material.

5. Conclusion

In this work, we introduce a modular framework for reconstructing complex 3D scenes from an image. We prioritize generalization by taking a divide-and-conquer approach rather than end-to-end. Our decomposing into multiple entities benefits from existing components, which effectively solve the established subtasks. We develop the necessary interfaces that enable the modules to function properly and finally yield a full 3D reconstruction. Our experiments decidedly show the advantage of the proposed method on various types of scenes. Considering the illustrated performance for diverse scenarios, we believe that our approach is a strong baseline and a stepping stone towards generalizable full 3D scene reconstruction from a single image.

Acknowledgement. This work was funded by the German Federal Ministry of Education and Research (BMBF), FKZ: 01IS22082 (IRRW). The authors are responsible for the content of this publication. The authors appreciate the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the NHR project b112dc IRRW. NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the German Research Foundation (DFG) – 440719683.



Input image | USL [22] | Ours (input view) | Ours (another view)

Figure 6. Qualitative results on the Hypersim dataset [63]. We compare against the USL method which also does not require 3D data for supervision. Our scenes are more complete, with significantly more objects being recognized and reconstructed.

Input image | InstPIFu [45] | DreamGaussian [69] | Ours

Figure 7. Qualitative results on real-world data. Animations are provided in the supplementary material video.

References

[1] 360cities.net. https://www.360cities.net/. Accessed: March 12, 2024. 1

[2] Kalyan Vasudev Alwala, Abhinav Gupta, and Shubham Tulsiani. Pre-train, self-train, distill: A simple recipe for supersizing 3d reconstruction. In *CVPR*, 2022. 1

[3] Aayush Bansal, Bryan Russell, and Abhinav Gupta. Marr revisited: 2d-3d alignment via surface normal prediction. In *CVPR*, 2016. 2

[4] Harry G Barrow, Jay M Tenenbaum, Robert C Bolles, and Helen C Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *Proceedings: Image Understanding Workshop*, 1977. 5

[5] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 3

[6] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 4

[7] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, 2018. 3

[8] Anpei Chen, Haofei Xu, Stefano Esposito, Siyu Tang, and Andreas Geiger. Lara: Efficient large-baseline radiance fields. In *ECCV*, 2024. 2, 3, 6

[9] Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *CVPR*, 2019. 2

[10] Yixin Chen, Junfeng Ni, Nan Jiang, Yaowei Zhang, Yixin Zhu, and Siyuan Huang. Single-view 3d scene reconstruction with high-fidelity shape and texture. In *International Conference on 3D Vision (3DV)*, 2024. 2

[11] Tao Chu, Pan Zhang, Qiong Liu, and Jiaqi Wang. Buol: A bottom-up framework with occupancy-aware lifting for panoptic 3d scene reconstruction from a single image. In *CVPR*, 2023. 2, 5, 6, 7

[12] Manuel Dahnert, Ji Hou, Matthias Nießner, and Angela Dai. Panoptic 3d scene reconstruction from a single rgb image. *NeurIPS*, 2021. 2

[13] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, 2023. 2, 4, 1

[14] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *NeurIPS*, 2024. 2, 4, 7

[15] Tien Do, Khiem Vuong, and Hyun Soo Park. Egocentric scene understanding via multimodal spatial rectifier. In *CVPR*, 2022. 2

[16] Yilun Du, Zhijian Liu, Hector Basevi, Ales Leonardis, Bill Freeman, Josh Tenenbaum, and Jiajun Wu. Learning to exploit stability for 3d scene parsing. *NeurIPS*, 2018. 2

[17] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM TOG*, 2020. 1, 2

[18] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 1981. 1

[19] 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 *ICCV*, 2021. 2, 5, 6, 7, 3, 4

[20] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *IJCV*, 2021. 2, 5, 3

[21] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 2013. 3

[22] Georgia Gkioxari, Nikhila Ravi, and Justin Johnson. Learning 3d object shape and layout without 3d supervision. In *CVPR*, 2022. 2, 5, 7, 8

[23] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 2

[24] Xian-Feng Han, Hamid Laga, and Bennamoun. Image-based 3d object reconstruction: State-of-the-art and trends in the deep learning era. *IEEE TPAMI*, 2019. 2

[25] Yannick Hold-Geoffroy, Kalyan Sunkavalli, Jonathan Eisenmann, Matt Fisher, Emiliano Gambaretto, Sunil Hadap, and Jean-François Lalonde. A perceptual measure for deep single image camera calibration. *CVPR*, 2018. 3

[26] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *arXiv preprint arXiv:2404.15506*, 2024. 7

[27] Siyuan Huang, Siyuan Qi, Yinxue Xiao, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. Cooperative holistic scene understanding: Unifying 3d object, layout, and camera pose estimation. *NeurIPS*, 2018. 2

[28] Siyuan Huang, Siyuan Qi, Yixin Zhu, Yinxue Xiao, Yuanlu Xu, and Song-Chun Zhu. Holistic 3d scene parsing and reconstruction from a single rgb image. In *ECCV*, 2018. 2

[29] Zixuan Huang, Stefan Stojanov, Anh Thai, Varun Jampani, and James M Rehg. Zeroshape: Regression-based zero-shot shape reconstruction. In *CVPR*, 2024. 7

[30] Hamid Izadinia, Qi Shan, and Steven M Seitz. Im2cad. In *CVPR*, 2017. 2

[31] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *CVPR*, 2023. 5, 2, 3

[32] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom

Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021. 2

[33] Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Matzen, Matthew Sticha, and David F. Fouhey. Perspective fields for single image camera calibration. In *CVPR*, 2023. 3, 5, 1, 8

[34] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024. 3, 5, 2

[35] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 2023. 4, 2

[36] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars. In *ECCV*, 2022. 1

[37] Muhammad Saif Ullah Khan, Alain Pagani, Marcus Liwicki, Didier Stricker, and Muhammad Zeshan Afzal. Three-dimensional reconstruction from a single rgb image using deep learning: A review. *Journal of Imaging*, 2022. 1

[38] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *ICCV*, 2023. 3, 2

[39] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM TOG*, 2017. 5

[40] Nilesh Kulkarni, Ishan Misra, Shubham Tulsiani, and Abhinav Gupta. 3d-relnet: Joint object and relational network for 3d prediction. In *ICCV*, 2019. 2

[41] Weicheng Kuo, Anelia Angelova, Tsung-Yi Lin, and Angela Dai. Mask2cad: 3d shape prediction by learning to segment and retrieve. In *ECCV*, 2020. 2

[42] Jinwoo Lee, Hyunsung Go, Hyunjoon Lee, Sunghyun Cho, Minhyuk Sung, and Junho Kim. CTRL-C: Camera calibration TRansformer with Line-Classification. In *ICCV*, 2021. 3

[43] Lin Li, Salman Khan, and Nick Barnes. Geometry to the rescue: 3d instance reconstruction from a cluttered scene. In *CVPRW*, 2020. 2

[44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2

[45] Haolin Liu, Yujian Zheng, Guanying Chen, Shuguang Cui, and Xiaoguang Han. Towards high-fidelity single-view holistic reconstruction of indoor scenes. In *ECCV*, 2022. 2, 5, 6, 7, 8, 3

[46] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. In *NeurIPS*, 2023. 2, 5, 3, 8

[47] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *CVPR*, 2024.

[48] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*, 2023. 2, 4, 1

[49] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. *SIGGRAPH Comput. Graph.*, 1987. 4, 5

[50] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *CVPR*, 2022. 3

[51] Alican Mertan, Damien Jade Duff, and Gozde Unal. Single image depth estimation: An overview. *Digital Signal Processing*, 2022. 3

[52] 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

[53] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM TOG*, 2022. 4

[54] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2

[55] 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 *CVPR*, 2020. 2

[56] Ege Ozguroglu, Ruoshi Liu, Dídac Surś, Dian Chen, Achal Dave, Pavel Tokmakov, and Carl Vondrick. pix2gestalt: Amodal segmentation by synthesizing wholes. *CVPR*, 2024. 7

[57] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 2

[58] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *ICLR*, 2023. 4

[59] Lu Qi, Jason Kuen, Yi Wang, Jiuxiang Gu, Hengshuang Zhao, Philip Torr, Zhe Lin, and Jiaya Jia. Open world entity segmentation. *TPAMI*, 2022. 3, 5

[60] Lu Qi, Jason Kuen, Tiancheng Shen, Jiuxiang Gu, Wenbo Li, Weidong Guo, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. High quality entity segmentation. In *ICCV*, 2023. 3, 2

[61] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *ECCV*, 2022. 1

[62] L.G. Roberts. *Machine Perception of Three-dimensional Solids*. M.I.T. Lincoln Laboratory, 1963. 2

[63] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *ICCV*, 2021. 7, 8

[64] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, 2022. 2, 3, 5, 1

[65] Taha Samavati and Mohsen Soryani. Deep learning-based 3d reconstruction: A survey. *Artificial Intelligence Review*, pages 1–45, 2023. 1

[66] Saurabh Saxena, Junhwa Hur, Charles Herrmann, Deqing Sun, and David J Fleet. Zero-shot metric depth with a field-of-view conditioned diffusion model. *arXiv preprint arXiv:2312.13252*, 2023. 7

[67] Alexander G Schwing, Sanja Fidler, Marc Pollefeys, and Raquel Urtasun. Box in the box: Joint 3d layout and object reasoning from single images. In *ICCV*, 2013. 2

[68] Karl Stelzner, Kristian Kersting, and Adam R Kosiorek. Decomposing 3d scenes into objects via unsupervised volume segmentation. *arXiv:2104.01148*, 2021. 2

[69] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *ICLR*, 2024. 2, 5, 6, 7, 8, 3, 4

[70] Antônio Wilson Vieira Thomas Lewiner, Hélio Lopes and Geovan Tavares. Efficient implementation of marching cubes' cases with topological guarantees. *Journal of Graphics Tools*, 8(2):1–15, 2003. 4, 5

[71] Edith Tretschk, Navami Kairanda, Mallikarjun BR, Rishabh Dabral, Adam Kortylewski, Bernhard Egger, Marc Habermann, Pascal Fua, Christian Theobalt, and Vladislav Golyanik. State of the art in dense monocular non-rigid 3d reconstruction. In *Comput. Graph. Forum*, 2023. 1

[72] Shubham Tulsiani, Saurabh Gupta, David F Fouhey, Alexei A Efros, and Jitendra Malik. Factoring shape, pose, and layout from the 2d image of a 3d scene. In *CVPR*, 2018. 2

[73] Stephen Tyree, Jonathan Tremblay, Thang To, Jia Cheng, Terry Mosier, Jeffrey Smith, and Stan Birchfield. 6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark. In *International Conference on Intelligent Robots and Systems (IROS)*, 2022. 5, 7, 3, 4

[74] Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, and Georgia Gkioxari. Multiview compressive coding for 3d reconstruction. In *CVPR*, 2023. 7

[75] Keyu Wu, Yifan Ye, Lingchen Yang, Hongbo Fu, Kun Zhou, and Youyi Zheng. Neuralhdhair: Automatic high-fidelity hair modeling from a single image using implicit neural representations. In *CVPR*, 2022. 1

[76] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *CVPR*, 2024. 3

[77] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. *CVPR*, 2023. 1, 2

[78] Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Amodal completion via progressive mixed context diffusion. In *CVPR*, 2024. 7

[79] Kai Yan, Fujun Luan, Miloš Hašan, Thibault Groueix, Valentin Deschaintre, and Shuang Zhao. Psdr-room: Single photo to scene using differentiable rendering. In *SIGGRAPH Asia*, 2023. 2

[80] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 3, 5, 2

[81] Yu-Ying Yeh, Zhengqin Li, Yannick Hold-Geoffroy, Rui Zhu, Zexiang Xu, Miloš Hašan, Kalyan Sunkavalli, and Manmohan Chandraker. Photoscene: Photorealistic material and lighting transfer for indoor scenes. In *CVPR*, 2022. 2

[82] Haobo Yuan, Xiangtai Li, Chong Zhou, Yining Li, Kai Chen, and Chen Change Loy. Open-vocabulary sam: Segment and recognize twenty-thousand classes interactively. In *ECCV*. Springer, 2024. 3, 5, 2

[83] Sergey Zakharov, Rares Andrei Ambrus, Vitor Campagnolo Guizilini, Dennis Park, Wadim Kehl, Fredo Durand, Joshua B Tenenbaum, Vincent Sitzmann, Jiajun Wu, and Adrien Gaidon. Single-shot scene reconstruction. In *5th Annual Conference on Robot Learning*, 2021. 2

[84] 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 *CVPR*, 2021. 2

[85] Xiang Zhang, Zeyuan Chen, Fangyin Wei, and Zhuowen Tu. Uni-3d: A universal model for panoptic 3d scene reconstruction. In *ICCV*, 2023. 2, 5, 6, 7

[86] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 2019. 2