OCH3R: Object-Centric Holistic 3D Reconstruction

Anonymous CVPR submission

Paper ID 14520

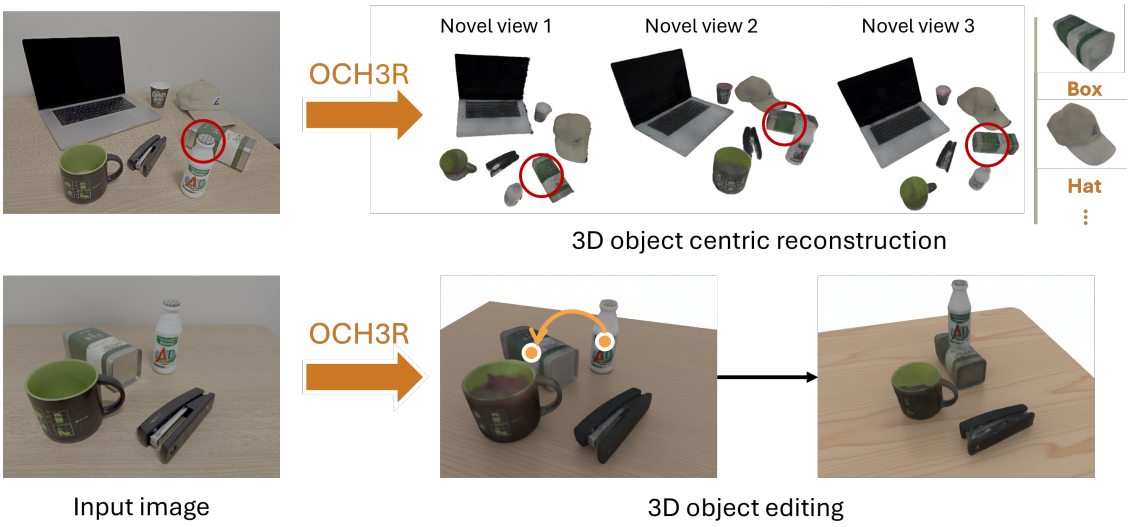

Figure 1. **OCH3R enables fully object-centric 3D scene reconstruction from a single RGB image.** Given one input view, OCH3R discovers all object instances, predicts their 6D poses, and reconstructs each object as a manipulable 3D Gaussian model in a single forward pass. Our feed-forward, per-pixel prediction framework supports selecting, moving, and rendering objects from arbitrary novel views without external segmentors or multi-stage pipelines. OCH3R produces amodally complete, editable 3D objects and generalizes to cluttered real scenes, enabling downstream tasks such as rearrangement and AR editing. The red circles highlight occluded regions that OCH3R successfully completes in 3D.

Abstract

Object-centric scene understanding is a fundamental challenge in computer vision. Existing approaches often rely on multi-stage pipelines that first apply pre-trained segmentors to extract individual objects, followed by per-object 3D reconstruction. Such methods are computationally expensive, fragile to segmentation errors, and scale poorly with scene complexity. We introduce ***OCH3R****, a unified framework for* ***O****bject-****C****entric* ***H****olistic* ***3****D* ***R****econstruction from a single RGB image. OCH3R performs one forward pass to simultaneously predict all object instances with their 6D poses and detailed 3D reconstructions. The key idea is a transformer architecture that predicts per-pixel attributes, including CLIP-based category embeddings, metric depth, normalized object coordinates (NOCS), and a fixed number of 3D Gaussians representing each object. To supervise these Gaussian reconstructions, we transform them into canonical space using the predicted 6D poses and align them with pre-rendered canonical ground truth, avoiding costly per-image Gaussian label generation. On standard indoor benchmarks, OCH3R achieves state-of-the-art performance across monocular depth estimation, open-vocabulary semantic segmentation, and RGB-only category-level 6D pose estimation, while producing high-fidelity, editable per-object reconstructions. Crucially, inference is fully feed-forward and scales independently of the number of objects, offering orders-of-magnitude speedups over conventional multi-stage pipelines in cluttered scenes.*

1. Introduction

Understanding a scene as a composition of discrete, posed objects from a single image is a long-standing goal in computer vision. Many downstream applications including robotic manipulation, AR editing, and simulation rely on

CVPR #14520

object-centric outputs [73], where each object is represented with geometry, pose, and semantics that can be selected or manipulated. We study the following problem: given a single RGB image of an indoor tabletop scene, recover all objects together with their 6D poses and corresponding 3D Gaussians in one forward pass.

Prior work largely falls into two groups. Scene-level, feed-forward methods (*e.g.*, one-pass Gaussian predictors) [5, 54, 68, 74, 82] are fast and photorealistic, but typically produce an undifferentiated "soup" of geometry without instance-level structure, canonical frames, or object poses. Object-level approaches [24, 25, 73] instead rely on multi-stage pipelines that begin with external open-vocabulary segmentors and then perform per-object reconstruction, alignment, and correction. These systems often depend on RGB-D inputs or category-specific priors, and they are fragile to upstream errors, difficult to scale with the number of objects, and not trained end-to-end. As a result, their robustness and accuracy degrade in cluttered tabletop scenes.

To address these limitations, we introduce **OCH3R**, a unified, object-centric, holistic 3D reconstructor that converts a single RGB image into a set of posed 3D objects in one pass. The key in our design is a 48-layer transformer that predicts dense, pixel-aligned attributes: CLIP [45]-based category embeddings, metric depth, normalized object coordinates (NOCS) [62], and a small set of 3D Gaussians [29] per pixel. During inference, object instances and their poses are recovered by clustering the semantic embeddings and estimating each object's $\mathrm{SIM}(3)$ pose using the predicted NOCS field.

To train the Gaussian representation, we allow Gaussians at each pixel to move freely off the pixel rays to compensate for (self-) occlusion. Rather than supervising Gaussians per training image [52, 53], we adopt Canonical-Space Supervision: per-object Gaussians are transformed into canonical space using its $\mathrm{SIM}(3)$ pose, and their renderings are optimized against pre-rendered ground truth in the canonical frame. This eliminates the need for costly per-image Gaussian labels and promotes amodal shape completion.

We train on a curated, large-scale dataset that integrates PACE [75], Omni6DPose [81], GSO [12], and Hypersim [48], offering broad coverage across object categories, poses, and occlusion patterns.

Our experiments show that OCH3R substantially outperforms previous baselines across all evaluated tabletop object-centric benchmarks. OCH3R delivers consistently higher geometric accuracy, better semantic alignment, and significantly more complete amodal reconstructions. Importantly, because OCH3R reconstructs all objects in a single forward pass, it achieves orders-of-magnitude faster inference than multi-stage pipelines while avoiding their brittleness to segmentation or pose-estimation errors. Together, these results highlight the effectiveness of our unified formulation and its practical advantages for real-world object-centric applications.

To summarize, our contributions are as follows:

1. We construct a large scale dataset for holistic object centric 3D scene representation. We assemble, relabel, and align PACE [75], Omni6DPose [81], GSO [12], and Hypersim [48] into a unified dataset designed for object centric 3D tasks, providing per instance masks, segmentation labels, $\mathrm{SIM}(3)$ poses, and 3D models.
2. We propose a model that yields high-fidelity 3D reconstructions, recovering fine-grained geometry and amodal structure, while jointly predicting semantics, monocular depth, and object poses in a single pass.
3. Experiments show that our model reconstructs real-world tabletop scenes with arbitrary numbers of objects, delivering more photorealistic and amodally complete results while running far faster than multi-stage pipelines.

2. Related Work

Feed-forward 3D reconstruction. Feed-forward 3D reconstruction maps one or a few images directly to a renderable 3D scene. Previous works have explored 3D representations including voxel grids [57, 58], multi-plane images [35, 59], meshes [17, 18], surfel [16], and radiance fields [21, 76]. More recently, 3D Gaussians [30] have emerged as a dominant representation for feed-forward regression thanks to their real-time differentiable rendering and compatibility with high-capacity 2D backbones.

Early feed-forward Gaussian predictors focus on single, centered objects, assigning one Gaussian to each input pixel and directly regressing its parameters without test-time optimization [52, 55, 71, 80]. These models deliver fast and high-quality reconstructions, but they assume clean, uncluttered inputs and cannot handle occlusions, multiple instances, or reassemble per-object predictions into a scene.

In contrast, scene-level Gaussian models predict a dense Gaussian field for an entire scene from one [53] or multiple images [5, 8, 65, 68, 74, 82]. Techniques including probabilistic splatting [5], cost-volume aggregation [8], depth conditioning [68], pose-free formulations [74] and large transformer backbones [82] have been explored to improve performance. While effective for novel-view synthesis, these scene-level approaches treat the world as a single undifferentiated cloud without instance decomposition, preventing downstream reasoning or interaction.

Object-centric scene reconstruction. A separate line of work performs object-centric scene reconstruction, explicitly recovering a set of 3D object instances and their layout. IM2CAD [26], Total3DUnderstanding [41], Zhang *et al*. [79], and CoReNet [44] reconstruct indoor scenes from

a single image by detecting furniture and room layout, then retrieving or predicting per-object geometry and enforcing consistency in a shared 3D frame. CAD- and RGB-D-based pipelines such as Mask2CAD [32], ROCA [19], CenterSnap [24], and ShAPO [25] further combine instance detection and depth with CAD retrieval or learned latent shape codes for each object, making them sensitive to upstream errors and computationally costly as the number of objects increases.

More recent methods introduce strong generative priors but largely retain this compositional, multi-stage design: Gen3DSR [1], CAST [73], and DepR [83] first apply monocular depth estimation and instance segmentation, then run object-level image-to-3D or diffusion models and compose the resulting objects into a coherent scene; MIDI [23] extends pre-trained image-to-3D generators to a multi-instance diffusion model that still takes segmented object crops as input. Consequently, computational cost and brittleness scale with the number and quality of segmented instances. With a sufficiently large dataset and a sufficiently powerful model, we show that single-view, object-aware 3D reconstruction can be approached as a direct, feed-forward prediction problem, rather than a fragile sequence of segmentation, retrieval, optimization, or generative refinement. In practice, this shift yields reconstructions that are not only orders-of-magnitude faster but also higher-fidelity.

3. Preliminaries

3D Gaussian Splatting. Gaussian Splatting [30] renders a scene represented by a finite set of anisotropic 3D Gaussian primitives by projecting each primitive to the image plane as a 2D Gaussian and alpha-compositing them in visibility order, yielding a fast, differentiable approximation to emission–absorption volume rendering. Compared with ray-sampled neural fields [40], splatting enables real-time rendering and efficient gradient backpropagation, and is widely used as the rendering backbone in recent feed-forward reconstruction methods [5, 6, 8, 52, 54, 55, 65, 71, 74, 82]; we adopt the same renderer throughout.

Normalized Object Coordinate Space. Normalized Object Coordinate Space (NOCS) [63] assigns each 3D point on an object instance a category-level, pose-invariant coordinate $\mathbf{c} \in [0,1]^3$ within a unit canonical cube whose axes are consistently aligned across instances of that category. We denote this unit cube as the *canonical space* and to its associated rigid coordinate system as the *canonical frame*.

Dense per-pixel NOCS predictions $\hat{\mathbf{c}}_{u,v}$ provide pixel-to-canonical correspondences that, together with the predicted 3D point map, are sufficient to recover an instance's category-level pose $\Pi = (s, R, \mathbf{t}) \in \mathrm{SIM}(3)$ (Sec. 4.2). By definition, Π transforms canonical coordinates to the camera (or scene) frame: $\mathbf{x}^{\text{cam}} = \Pi(\mathbf{x}^{\text{can}}) = sR\mathbf{x}^{\text{can}} + \mathbf{t}$, with inverse mapping given by $\Pi^{-1}(\mathbf{x}^{\text{cam}}) = s^{-1}R^{\top}\cdot(\mathbf{x}^{\text{cam}} - \mathbf{t})$.

4. Method

4.1. Problem formulation and notation

Given a single RGB image $I \in \mathbb{R}^{H\times W\times 3}$ with unknown intrinsics K, OCH3R converts the image into an object-centric scene: a set of instances, each with a category-level $\mathrm{SIM}(3)$ pose and a high-fidelity 3D representation. Achieving this in one pass requires pixel-aligned predictions that are sufficient to (i) discover instances and semantics, (ii) recover a metric similarity transform for each object, and (iii) assemble amodally complete Gaussian representation of each object into an interactive scene.

Specifically, for each pixel (u, v), our network Φ outputs:

$$\Phi(I)_{u,v} = (\hat{\mathbf{e}}_{u,v}, \hat{d}_{u,v}, \hat{\mathbf{c}}_{u,v}, \hat{\mathcal{G}}_{u,v}), \tag{1}$$

where $\hat{\mathbf{e}}_{u,v} \in \mathbb{R}^{512}$ is the semantic label of the object that this pixel belongs to. We define semantic label of an object as the CLIP [45] embedding of the object's category name [34]. $\hat{d}_{u,v} \in \mathbb{R}^{+}$ is the predicted metric depth. It enables back-projecting the pixel into 3D space via $\hat{\mathbf{p}}_{u,v} = \hat{d}_{u,v} \cdot K^{-1}\begin{bmatrix} u & v & 1\end{bmatrix}^{\top} \in \mathbb{R}^3$. $\hat{\mathbf{c}}_{u,v} \in [0,1]^3$ is the predicted NOCS [63] coordinate, which enables $\mathrm{SIM}(3)$ pose recovery.

$\hat{\mathcal{G}}_{u,v} = \{g^{(i)}_{u,v}\}_{i=1}^{k}$ is a small set of anisotropic 3D Gaussian primitives (we use $k = 2$) that will be aggregated into per-object reconstruction:

$$g^{(i)}_{u,v} = (\boldsymbol{\mu}^{(i)}_{u,v}, \Sigma^{(i)}_{u,v}, \alpha^{(i)}_{u,v}, \mathbf{S}^{(i)}_{u,v}), \tag{2}$$

with mean $\boldsymbol{\mu}^{(i)}_{u,v} \in \mathbb{R}^3$, covariance $\Sigma^{(i)}_{u,v} \in \mathbb{S}^3_{++}$, opacity $\alpha^{(i)}_{u,v} \in (0,1)$, and RGB spherical harmonics (SH) coefficients $\mathbf{S}^{(i)}_{u,v} \in \mathbb{R}^{3(L+1)^2}$ (order L). The Gaussian parameters are defined and predicted in camera frame, and will be transformed into each object's canonical frame for supervision and inference (Sec. 4.3).

Following VGGT [64], we also predict the camera field of view $(\hat{\theta}_w, \hat{\theta}_h)$ of the input image and construct K with $f_w = \frac{W}{2\tan(\hat{\theta}_w/2)}$, $f_h = \frac{H}{2\tan(\hat{\theta}_h/2)}$ and principal point at image center.

In Sec. 4.2, we show how $\hat{\mathbf{e}}$, $\hat{d}$, $\hat{\mathbf{c}}$, and $\hat{\mathcal{G}}$ are used to discover instances, estimate object poses, and assemble reconstructed objects. Sec. 4.3 introduces Canonical Space Supervision (CSS) that trains Gaussians to be object-aligned and amodally complete without per-image Gaussian labels. Sec. 4.4 summarizes architectural and training details. Our full pipeline is given in Fig. 2.

4.2. Assembling objects from dense predictions

Instance discovery. At inference time, for each pixel, we first compute the cosine similarity between the predicted embedding $\hat{\mathbf{e}}_{u,v}$ and a set of predefined category

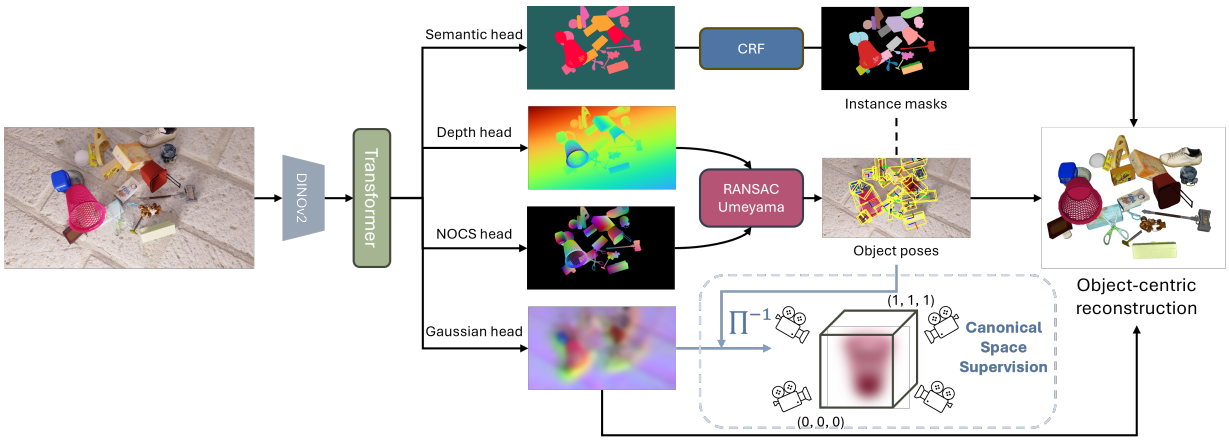

Figure 2. **Overview of our single-view object-centric 3D reconstruction pipeline.** Given a single RGB input, we extract dense DINOv2 features and feed them to a transformer that predicts per-pixel depth, CLIP-space semantic embeddings, NOCS coordinates, and Gaussian primitives. A CRF refines semantic affinities to produce coherent instance masks. For each instance, we estimate a category-level SIM(3) pose via RANSAC-Umeyama using the predicted NOCS-to-3D correspondences, enabling a transformation from camera space into the canonical object frame. The per-pixel Gaussians are then grouped and transformed into canonical space, where Canonical-Space Supervision (CSS) trains them to form amodally complete, compact 3D Gaussians. Aggregating all reconstructed objects yields an interactive, object-aligned scene representation from a single image.

name CLIP embeddings $\{\mathbf{l}_c\}$. We then apply a fully connected conditional random field (CRF) [31], using unary potentials defined as $-\log \frac{\exp(\cos(\hat{\mathbf{e}}_{u,v}, \mathbf{l}_c)/\tau)}{\sum_{c'} \exp(\cos(\hat{\mathbf{e}}_{u,v}, \mathbf{l}_{c'})/\tau)}$ for each category c, where τ denotes the temperature parameter in the softmax function. Pairwise potentials are defined as $\cos(\hat{\mathbf{e}}_{u,v}, \hat{\mathbf{e}}_{u',v'})$. For more details about CRF, we refer the reader to [31]. This process yields groups $\{\hat{\mathcal{P}}_j\}$, where each $\hat{\mathcal{P}}_j$ represents the set of pixels corresponding to object j.

Pose estimation. With the pixels for each object instance identified, we use their predicted NOCS [62] coordinates $\mathbf{c}_{u,v}$ to determine the object's precise SIM(3) pose in the scene. The NOCS coordinates establish a correspondence between a point's observed position in the scene and its standardized position within a unit canonical cube. We can therefore solve for the similarity transformation $\hat{\Pi}_j = (\hat{s}_j, \hat{R}_j, \hat{\mathbf{t}}_j)$, representing the scale, rotation, and translation, which maps the canonical space of object j to the camera space. This is achieved by minimizing the alignment error between the back-projected 3D points and the transformed NOCS coordinates over all pixels belonging to that instance:

$$\hat{\Pi}_j = \arg\min_{\Pi} \sum_{(u,v)\in\hat{\mathcal{P}}_j} \left\| \hat{\mathbf{p}}_{u,v} - \Pi(\hat{\mathbf{c}}_{u,v}) \right\|^2, \qquad (3)$$

where $\Pi(\hat{\mathbf{c}}_{u,v}) = sR \cdot \hat{\mathbf{c}}_{u,v} + \mathbf{t}$. This optimization can be solved using Umeyama algorithm [61] with RANSAC [15]. The resulting inverse transformation, $\hat{\Pi}_j^{-1}$, gives us a direct mapping from the cluttered scene into the clean canonical space for each object. Notably, this NOCS prediction can also be used to differentiate object instances with the same category name but that are adjacent in the mask, where CRF alone may not be enough. We run multiple RANSACs within each CRF-generated mask, and output objects when there are still enough inliers.

Object Gaussians. With the instance mask and estimated pose in hand, we obtain each object's canonical-space Gaussian representation by transforming every predicted Gaussian mean as $\boldsymbol{\mu}_{u,v}^{\text{can},(i)} = \hat{\Pi}_j^{-1}(\boldsymbol{\mu}_{u,v}^{(i)})$ for all $(u,v) \in \hat{\mathcal{P}}_j, i \in \{1,\ldots,k\}$. The resulting set of transformed Gaussians forms the complete canonical representation of object j.

Efficiency. Since OCH3R predicts all per-pixel quantities in one forward pass, every object is reconstructed at once. Our CUDA CRF runs in roughly 200 ms per image, and our CUDA RANSAC Umeyama adds under 10 ms per object, making its cost negligible. Consequently, runtime is nearly invariant to scene complexity and remains far below prior pipelines [56, 73, 83], which synthesize each object through iterative diffusion denoising and often require relation-graph optimization that grows quadratically with the number of objects.

4.3. Canonical-Space Supervision

One key challenge for our Gaussian prediction network is that it must infer a full, amodal set of object Gaussians from only the pixels that are actually visible. A natural idea is to use pre-optimized object Gaussians and place them in the camera frame so they can serve as ground-truth supervision. However, there lacks one-to-one correspondence between visible pixels and ground-truth object Gaussians. To address this, we introduce *Canonical Space Supervision* (Fig. 3), a strategy that transfers training signals to the ob-

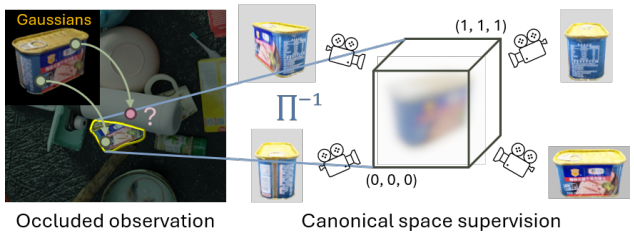

Figure 3. **Canonical Space Supervision (CSS).** Predicted per-pixel Gaussians are transformed into the object's canonical frame via the ground-truth pose Π^{-1}. In canonical space, they are supervised against pre-rendered multi-view ground-truth images, providing clean amodal signals that resolve occlusions and enforce compact, object-aligned Gaussian reconstructions.

ject's canonical frame, where clean targets are available.

Concretely, we place each training object mesh in the canonical frame, and pre-render a set of N views $\mathcal{V} = \{I_n^{\text{gt}}\}_{n=1}^{N}$, where $I_n^{\text{gt}} \in \mathbb{R}^{H_{\text{can}} \times W_{\text{can}} \times 3}$. We set $N = 42$, $H_{\text{can}} = W_{\text{can}} = 512$. This is done once per object and reused across all images containing that object.

During training, we use ground truth object masks $\{\mathcal{P}_j\}$ to extract pixels of each object j. We transform the predicted Gaussians per object from the camara space into the canonical space with the ground truth object pose.

With the transformed Gaussians $\{g_{u,v}^{\text{can},(i)}\}$ of the object, we can render images of that object in its canonical space using the same camera angles as $\mathcal{V}$ with a differentiable Gaussian rasterizer [29]. Let $\{\hat{I}_n\}_{n=1}^{N}$ denote the rendered images. Then the CSS loss is calculated by

$$\mathcal{L}_{\text{CSS}} = \sum_{n=1}^{N} \Big(\|I_n^{\text{gt}} - \hat{I}_n\|_1 + \lambda_{\text{SSIM}}(1 - \text{SSIM}(I_n^{\text{gt}}, \hat{I}_n)) \Big). \tag{4}$$

Occlusion handling via off-ray offsets. Each Gaussian mean is anchored at $\hat{\mathbf{p}}_{u,v}$ with a predicted camera-space offset $\mathbf{\Delta}_{u,v}^{(i)}$, allowing a visible pixel to spawn Gaussians behind the first surface. Because CSS supervises in canonical space, occluded Gaussians receive gradients even when not visible in the input. Optionally, we regularize with an annealed small-offset prior: $\mathcal{L}_{\text{reg}} = \sum_{(u,v)\in\mathcal{P}_j} \sum_{i=1}^{k} \text{ReLU}(\|\mathbf{\Delta}_{u,v}^{(i)} - \tau_{\text{offset}}\|_1)$.

4.4. Architecture and training details

Architecture. Inspired by VGGT [64], OCH3R uses a DINOv2 backbone [42] followed by a 48-layer ViT encoder and DPT-style decoder heads [46]. The input image is patchified by DINOv2 and processed by global self-attention through the encoder.

For dense tasks, each head takes features from four intermediate encoder layers (lateral skips), projects them to a common width, and upsamples to the image resolution with convolutional fusion.

For camera FOV prediction, we append a learnable camera token that is updated by a small stack of transformer blocks with adaptive layer-norm modulation and iterative refinement; a linear layer regresses the field-of-view angles $(\hat{\theta}_w, \hat{\theta}_h)$.[1]

For Gaussian prediction we decouple geometry and appearance. A geometry head outputs per-pixel off-ray offsets $\mathbf{\Delta}_{u,v}^{(i)}$ (for i=1, . . . , k), which are added to the back-projected point $\hat{\mathbf{p}}_{u,v}$ to obtain camera-frame means $\boldsymbol{\mu}_{u,v}^{(i)}$. An appearance/shape head predicts canonical-frame scales $\boldsymbol{\sigma}_{u,v}^{(i)}$, unit quaternions $\mathbf{q}_{u,v}^{\text{can},(i)}$, opacities $\alpha_{u,v}^{(i)}$, and SH coefficients $\mathbf{S}_{u,v}^{(i)}$. Following NoPoSplat [74], we provide an *RGB shortcut* to the appearance/shape head to improve fine texture details in 3D reconstruction. All parameters of each Gaussian and the k Gaussians of every pixel are simply concatenated.

Training. We optimize all tasks jointly with AdamW and a cosine learning-rate schedule. DINOv2 is initialized from public weights and fine-tuned end-to-end; the full list of hyperparameters is provided in the appendix.

Depth. We supervise canonical inverse depth as in Depth Pro [3]. Let f_w be the horizontal focal length (pixels) and W the image width, define $C = \frac{f_w}{W \cdot d}$. Our model outputs $\hat{C}$ and is trained to minimize $\mathcal{L}_{\text{depth}} = \|\hat{C} - C\|_2 + \lambda_\nabla \Big(\|\nabla_x(\hat{C} - C)\|_2 + \|\nabla_y(\hat{C} - C)\|_2 \Big)$. At test time, we recover metric depth by $\hat{d}_{u,v} := \frac{\hat{f}_{\text{px}}}{W \cdot \hat{C}_{u,v}}$.

Semantics. During training time, we dynamically compute cosine similarities between the embedding of pixel and all the words that appear in the training image. We then encourage the embedding of the pixel to align with the ground-truth class by using the computed cosine similarities as the logits for softmax, with cross entropy loss.

NOCS. We reformulate NOCS coordinate regression as a bin classification task augmented with a learnable offset, which implicitly resolves ambiguities in symmetric objects. Each axis in NOCS (i.e., *xyz*) is discretized into M=64 centered bins. We supervise the bin classification using a cross-entropy loss and the offset prediction using a mean squared error loss. The total loss is averaged over all foreground object pixels.

Gaussians (CSS). Gaussian supervision follows Canonical Space Supervision discussed in Sec. 4.3.

Camera FOV. We supervise $(\hat{\theta}_w, \hat{\theta}_h)$ with a robust Huber loss on angles: $\mathcal{L}_{\text{cam}} = \|\hat{\theta}_w - \theta_w\|_\epsilon + \|\hat{\theta}_h - \theta_h\|_\epsilon$.

Task losses are combined with homoscedastic uncertainty weighting [28]; ablations are in the appendix.

[1] Our implementation retains VGGT's iterative refinement; translation/rotation channels are present in the token state but only FOV is used at test time.

5. Dataset

Existing indoor benchmarks for monocular depth estimation and open vocabulary semantic segmentation [10, 50, 51] primarily emphasize room layout and large furniture, while providing limited coverage of the object interaction scale that is central to everyday visual tasks. In contrast, progress in embodied perception [4, 14, 49, 66] and in mobile AR or MR systems [13, 20] requires accurate modeling of small, manipulable, and semantically diverse objects such as cups, tools, and containers that humans routinely interact with.

To support this direction, we construct a new evaluation benchmark by integrating several real world datasets tailored to this domain. Specifically, we include the validation split of HOPE [60] and the test splits of YCB Video [67], PACE [75], Omni6DPose [81] (OMNI), and NOCS [63]. For training, we curate and align four large scale sources: the training splits of PACE and Omni6DPose, Google Scanned Objects [12] renderings from FoundationPose [66], and HyperSim [48].

6. Experiments

We first evaluate holistic 3D object-centric reconstruction from a single RGB image in Sec. 6.1. Sec. 6.2 then demonstrates that beyond 3D reconstruction, our method also delivers state-of-the-art zero-shot performance on depth estimation, segmentation, and object pose prediction. Sec. 6.3 examines key design choices through targeted ablations.

6.1. 3D Reconstruction

We evaluate OCH3R against Gen3DSR [1], ACDC [11], and a unified glued baseline that uses instance masks from SAM2 [47] and GroundingDINO [39], object poses from MonoDiff9D [38], and depth from DepthPro [3]. We denote this pipeline as AoE (Army of Experts). Since the baselines are extremely slow, we randomly sample ten images from each dataset in our benchmark.

Following prior work [1, 73], we report chamfer distance and F-1@0.1 between the predicted and ground truth meshes, and CLIP similarity between the rendered and ground truth images. All backgrounds are manually normalized to white for both predictions and ground truth.

As shown in Tab. 1, our method establishes a clear margin over all baselines. On PACE, OCH3R reduces the Chamfer Distance from 0.31 (Gen3DSR) and 0.35 (AoE) to 0.18, while more than doubling the best F-1 score (45.00 versus AoE's 21.39). Similar trends hold across all remaining datasets: on YCB-V, OCH3R achieves 0.17 CD (a 26 percent improvement over AoE and a 48 percent improvement over Gen3DSR) and reaches 22.71 F-1, surpassing the strongest baseline by over 10 points. On HOPE, OCH3R attains 83.69 CLIP similarity, exceeding Gen3DSR by +6.1 and AoE by +19.9. For NOCS real, OCH3R's gains are the most pronounced, improving CD from 0.15 to 0.07 and F-1 from 38.01 to 76.77. These accuracy improvements come alongside a dramatic speedup: our 0.7 s inference time per image is roughly 2,000x faster than Gen3DSR (25.6 min) and ACDC (22.1 min), while also running more than 30x faster than AoE (21.6 s). It demonstrates the advantage of our unified per pixel prediction formulation. Some qualitative results are shown in Fig. 4.

6.2. Individual task performance

Zero-shot metric depth. Accurate metric depth from a single RGB image is essential to our pipeline, as it defines the anchor positions for our 3D Gaussians. We evaluate OCH3R on five datasets against seven state-of-the-art baselines using three standard metrics: δ_1 [33], AbsRel, and RMSE. Additional metrics (δ_2, δ_3, log10, $\mathrm{RMSE}_{\log}$, SI-log) are provided in the Supplementary.

As shown in Tab. 2, OCH3R achieves leading performance on all three metrics for PACE, HOPE, and NOCS-real, and on δ_1 for YCB-V. Although Metric3D V2 [22] and Depth Pro [3] perform slightly better on OMNI, the margin is minimal; moreover, Metric3D V2 requires ground-truth camera intrinsics, giving it an inherent advantage, yet it is still surpassed by OCH3R on four of the five datasets. Depth Anything V2 [72], trained primarily for relative depth and fine-tuned for metric estimation, shows high domain sensitivity, excelling on YCB-V (narrowly ahead of OCH3R) but degrading substantially elsewhere. Overall, OCH3R attains the best results in 10 of 15 metric–dataset combinations and delivers the strongest average performance across benchmarks.

Zero-shot semantic segmentation. Open-vocabulary semantic segmentation assigns per-pixel labels drawn from a potentially open set of natural-language concepts. To build an evaluation vocabulary disjoint from training, we aggregate names from the test datasets, common indoor categories from ADE20K [84, 85], and additional household items.

We report standard OVSS metrics: mIoU and FB-IoU. Given the difficulty of segmenting fine-grained, cluttered indoor scenes, we additionally report hit@5, which allows each method to produce up to five candidate labels per pixel. Since FB-IoU already captures the ability to separate foreground from background, the remaining metrics are computed on foreground regions only.

Tab. 3 compares OCH3R with seven OVSS baselines. Across five datasets and three metrics (15 settings), OCH3R ranks first in 12 and achieves the best average rank (1.27). It leads all three metrics on PACE, YCB-V, and NOCS-Real; mIoU and FB-IoU on HOPE; and FB-IoU on OMNI. Averaged across datasets, OCH3R obtains 11.04 mIoU, 83.18 FB-IoU, and 83.56 hit@5, outperforming the strongest

CVPR #14520

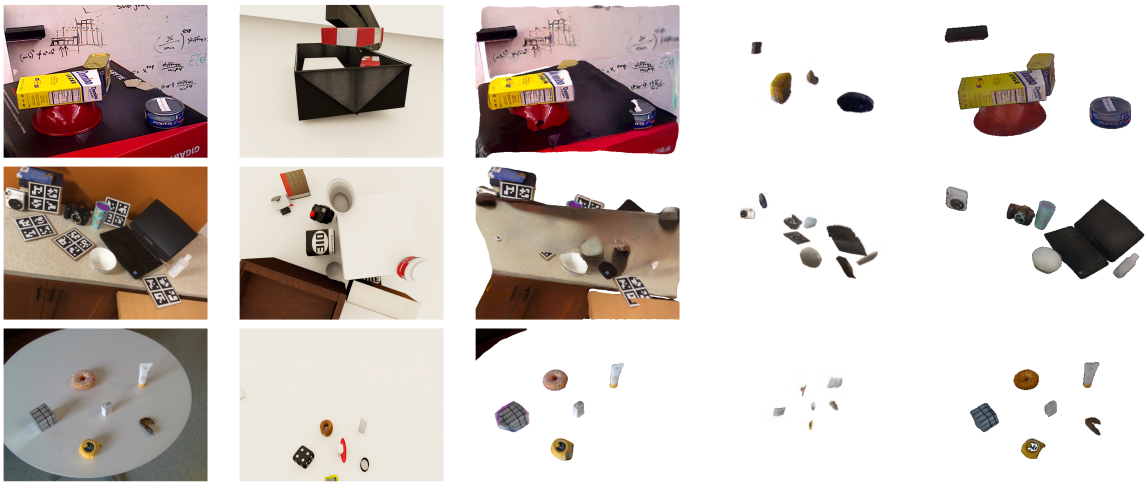

Figure 4. **Qualitative comparison of single-image 3D object-centric reconstruction.** Given a single RGB input, we compare our method (OCH3R) with ACDC, Gen3DSR, and AoE (Army of Experts: SAM2 + GroundingDINO + MonoDiff9D + DepthPro). Prior methods often yield incomplete geometry, distorted textures, or missing objects. OCH3R reconstructs sharper, more complete, and semantically consistent objects across diverse scenes.

Model	PACE			OMNI			YCB-V			HOPE			NOCS real			Time per image↓
	CD↓	F-1↑	CLIP↑	CD↓	F-1↑	CLIP↑	CD↓	F-1↑	CLIP↑	CD↓	F-1↑	CLIP↑	CD↓	F-1↑	CLIP↑	
ACDC [11]	0.69	0.00	73.74	–	–	72.23	0.32	7.76	60.15	1.42	1.92	63.75	3.44	1.72	69.74	22.1 min
Gen3DSR [1]	0.31	12.93	81.61	–	–	82.00	0.33	8.85	75.26	0.35	12.05	77.63	0.26	17.73	75.33	25.6 min
AoE	0.35	21.39	77.97	–	–	69.31	0.23	12.45	66.05	0.18	**47.71**	63.80	0.15	38.01	70.62	21.6 s
OCH3R (ours)	**0.18**	**45.00**	**83.15**	–	–	**82.78**	**0.17**	**22.71**	**78.80**	**0.14**	40.08	**83.69**	**0.07**	**76.77**	**85.90**	**0.7 s**

Table 1. Comparison of 3D reconstruction and semantic consistency across datasets using CD (Chamfer Distance, lower is better), F-1 score, and CLIP similarity.

baseline by +2.44 mIoU (MAFT+ [27]), +36.80 FB-IoU (SAN [70]), and +2.47 hit@5 (MAFT+).

Zero-shot pose estimation. We further evaluate OCH3R 's ability to recover category-level 6D object poses from a single RGB image in a zero-shot setting. Due to space constraints, the full quantitative comparison with AG-Pose [37], SecondPose [7], and MonoDiff9D [38] across five indoor benchmarks is provided in the Supplementary. AG-Pose and SecondPose require RGB-D inputs, so we supply them with our predicted depths. Following MonoDiff9D, we report accuracy within 10 cm, within 10°, and under the joint 10°/10 cm criterion.

Across all datasets, OCH3R shows consistent gains on the stricter angular and joint metrics. It attains the highest 10° and joint accuracy on PACE and HOPE and improves the joint metric on NOCS-real. For example, on PACE, OCH3R improves the 10° rate from 15.1 to 25.9 and the joint 10°/10 cm rate from 8.6 to 14.0 compared to AG-Pose. These results indicate that the unified 3D representation learned by OCH3R naturally supports precise, canonically aligned object poses without any dataset-specific fine-tuning.

In summary, our model's strong performance across monocular depth estimation, open-vocabulary semantic segmentation, and pose estimation jointly enables state-of-the-art 3D reconstruction quality, producing geometrically precise, semantically coherent, and canonically aligned scene representations.

6.3. Ablation

Multi-task learning. To demonstrate the advantage of using a unified model for multiple traditionally separated tasks, we retrain the model while removing one head at a time. Experiments shows that dropping semantic embeddings causes large pose and Gaussian degradations and also hurts depth. More broadly, removing any head weakens the remaining tasks, and the full four-head variant performs best. Full per-dataset quantitative results are provided in the Supplementary.

Offset in camera space vs. directly in canonical space. We also tried predicting Gaussian parameters directly in the canonical frame, bypassing the offset-along-ray formulation. As shown in Fig. 5, removing the geometric scaffold of camera rays causes the network to collapse into

Model	PACE			OMNI			YCB-V			HOPE			NOCS real		
	$\delta_1\uparrow$	AbsRel↓	RMSE↓	$\delta_1\uparrow$	AbsRel↓	RMSE↓	$\delta_1\uparrow$	AbsRel↓	RMSE↓	$\delta_1\uparrow$	AbsRel↓	RMSE↓	$\delta_1\uparrow$	AbsRel↓	RMSE↓
NeWCRFs [78]	17.00	0.7356	0.5308	21.41	0.6432	0.9934	52.71	0.2740	0.2699	5.37	0.8073	0.5029	36.00	0.3352	0.3586
ZoeDepth [2]	8.20	0.8508	0.6146	20.73	0.5794	0.9406	52.84	0.2704	0.2750	0.53	0.9814	0.6055	34.08	0.4399	0.5275
Metric3D V2 [22]	1.32	0.4695	0.3995	**73.73**	**0.1788**	0.7963	15.44	0.2791	0.2538	0.01	0.4039	0.2968	8.50	0.3273	0.3571
Depth Anything V2 [72]	13.07	0.7373	0.5906	5.83	0.9009	1.1728	50.74	0.1899	**0.2029**	0.07	1.0364	0.6518	2.44	0.5507	0.6088
Depth Pro [3]	33.63	0.5603	0.5610	57.37	0.2742	**0.7940**	49.04	0.2078	0.2148	23.09	0.4035	0.3013	81.58	0.1370	0.1831
VGGT [64]	55.58	0.2415	0.2096	57.14	0.2272	1.0365	63.88	**0.1822**	0.2131	37.17	0.4313	0.2844	92.26	0.0965	0.1394
UniDepth V2 [43]	35.05	0.5076	0.4820	22.80	0.7221	1.1007	32.70	0.2548	0.2499	8.63	0.7622	0.6830	78.92	0.1778	0.2366
OCH3R (ours)	**94.82**	**0.0834**	**0.1039**	42.77	0.2900	0.8042	**69.96**	0.1933	0.2044	**61.36**	**0.2192**	**0.1812**	**98.43**	**0.0923**	**0.1066**

Table 2. **Monocular metric depth estimation results** on PACE, OMNI, YCB-V, HOPE, and NOCS real. Each dataset block reports δ_1 (in percentage), AbsRel, and RMSE. Bold indicates the best result, and underline indicates the second best.

Model	PACE			OMNI			YCB-V			HOPE			NOCS real		
	mIoU↑	FB-IoU↑	hit@5↑	mIoU↑	FB-IoU↑	hit@5↑	mIoU↑	FB-IoU↑	hit@5↑	mIoU↑	FB-IoU↑	hit@5↑	mIoU↑	FB-IoU↑	hit@5↑
LSeg [34]	0.26	17.23	8.52	2.43	4.55	16.31	2.63	12.41	23.05	0.30	12.09	20.69	5.48	8.73	79.82
OVSeg [36]	1.28	13.57	28.84	11.24	2.94	52.19	1.29	12.20	58.66	0.31	11.87	25.99	1.87	8.54	79.21
ODISE [69]	1.69	13.70	36.05	13.27	3.66	58.54	4.51	12.17	76.76	0.99	11.87	68.77	6.31	8.53	90.60
FC-CLIP [77]	3.67	13.56	61.62	22.29	3.61	71.11	3.69	12.15	69.96	1.85	11.91	71.80	4.72	9.40	94.68
CAT-Seg [9]	1.26	21.48	30.53	9.61	4.14	52.62	2.38	12.12	65.78	2.87	11.91	75.97	5.31	8.79	95.80
SAN [70]	1.70	71.05	31.87	13.75	40.82	53.09	4.25	44.77	72.83	3.27	29.96	63.94	5.22	45.30	94.56
MAFT+ [27]	3.66	13.76	64.39	**23.02**	3.61	**76.69**	5.00	12.16	79.01	3.21	11.87	**89.23**	8.10	8.53	96.14
OCH3R (ours)	**7.34**	**94.95**	**84.76**	18.43	**84.21**	75.23	**9.14**	**72.00**	**80.03**	**6.90**	**71.78**	79.27	**13.40**	**92.95**	**98.51**

Table 3. **Open-vocabulary semantic segmentation results** on PACE, OMNI, YCB-V, HOPE, and NOCS real. Each dataset block reports mIoU, FB-IoU, and hit@5 in percentages. Bold indicates the best result, and underline indicates the second best.

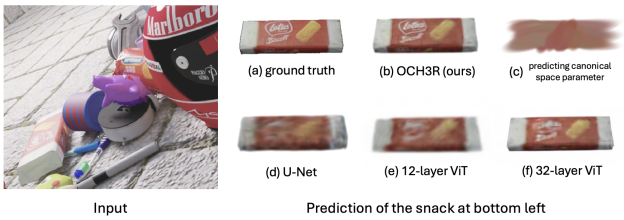

Figure 5. Qualitative ablations showing that (i) predicting Gaussians directly in canonical space collapses, and (ii) OCH3R's formulation remains robust across model scales and architectures.

an unstructured blob, confirming that unconstrained pixel-to-Gaussian mapping is highly underdetermined. By instead predicting per-pixel offsets in camera space and using Canonical-Space Supervision to organize the final layout, OCH3R obtains a stable and expressive inductive bias that enables high-quality reconstruction.

Model scale and architecture. The core of our method, particularly the multi-task learning paradigm and Canonical-Space Supervision, is orthogonal to model architecture, so in principle any dense predictor could be used. We validate this by testing a U-Net (matched in parameter count to the 32-layer ViT) and two smaller ViT variants. As shown in Fig. 5, while larger models yield sharper textures and cleaner geometry, all architectures successfully recover the object's overall shape. This confirms that OCH3R's pipeline transfers across dense predictors and scales well with model capacity.

7. Conclusion

In this paper, we try to address the long standing problem of understanding a scene as a set of discrete, posed objects from a single RGB image. Rather than relying on multi stage pipelines that decompose the task into segmentation, per object reconstruction, and post hoc alignment, we proposed OCH3R, a unified, feed forward model that predicts all object instances, their category level $\mathrm{SIM}(3)$ poses, and high fidelity 3D Gaussians in a single pass. The key ingredients are a transformer that produces dense, pixel aligned attributes (metric depth, CLIP based semantics, NOCS coordinates, and per pixel Gaussians), a simple inference procedure for instance discovery and pose estimation, and canonical space supervision that trains amodally complete Gaussians without per image Gaussian labels.

To support this setting, we assembled a large scale dataset for holistic object centric 3D scene representation by aligning PACE, Omni6DPose, HOPE, YCB-Video and NOCS into a unified benchmark with per instance masks, semantics, 6D poses, and 3D models. Across this benchmark, OCH3R outperforms previous baselines on monocular depth estimation, open vocabulary segmentation, and category level pose prediction, while producing more complete, editable 3D object reconstructions with feed forward inference that scales essentially independently of the number of objects.

References

[1] Andreea Ardelean, Mert Özer, and Bernhard Egger. Generalizable 3d scene reconstruction via divide and conquer from a single view. In *International Conference on 3D Vision (3DV)*, pages 616–626, 2025. 3, 6, 7

[2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023. 8

[3] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second, 2025. 5, 6, 8

[4] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023. 6

[5] David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction, 2024. 2, 3

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

[7] Yamei Chen, Yan Di, Guangyao Zhai, Fabian Manhardt, Chenyangguang Zhang, Ruida Zhang, Federico Tombari, Nassir Navab, and Benjamin Busam. Secondpose: Se(3)-consistent dual-stream feature fusion for category-level pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9959–9969, 2024. 7

[8] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. *MVSplat: Efficient 3D Gaussian Splatting from Sparse Multi-view Images*, page 370–386. Springer Nature Switzerland, 2024. 2, 3

[9] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation, 2024. 8

[10] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes, 2017. 6

[11] Tianyuan Dai, Josiah Wong, Yunfan Jiang, Chen Wang, Cem Gokmen, Ruohan Zhang, Jiajun Wu, and Li Fei-Fei. Automated creation of digital cousins for robust policy learning. In *Conference on Robot Learning (CoRL)*, 2024. 6, 7

[12] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B. McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items, 2022. 2, 6

[13] Ruofei Du, Eric Turner, Maksym Dzitsiuk, Luca Prasso, Ivo Duarte, Jason Dourgarian, Joao Afonso, Jose Pascoal, Josh Gladstone, Nuno Cruces, Shahram Izadi, Adarsh Kowdle, Konstantine Tsotsos, and David Kim. Depthlab: Real-time 3d interaction with depth maps for mobile augmented reality. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, page 829–843, New York, NY, USA, 2020. Association for Computing Machinery. 6

[14] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[15] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981. 4

[16] Yiming Gao, Yan-Pei Cao, and Ying Shan. Surfelnerf: Neural surfel radiance fields for online photorealistic reconstruction of indoor scenes, 2023. 2

[17] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn, 2020. 2

[18] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan C. Russell, and Mathieu Aubry. Atlasnet: A papier-mâché approach to learning 3d surface generation, 2018. 2

[19] 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 (CVPR)*, pages 4022–4031, 2022. 3

[20] Aleksander Holynski and Johannes Kopf. Fast depth densification for occlusion-aware augmented reality. *ACM Transactions on Graphics (TOG)*, 37(6), 2018. 6

[21] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d, 2024. 2

[22] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10579–10596, 2024. 6, 8

[23] Zehuan Huang, Yuan-Chen Guo, Xingqiao An, Yunhan Yang, Yangguang Li, Zi-Xin Zou, Ding Liang, Xihui Liu, Yan-Pei Cao, and Lu Sheng. Midi: Multi-instance diffusion for single image to 3d scene generation, 2025. 3

[24] 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, 2022. 2, 3

[25] Muhammad Zubair Irshad, Sergey Zakharov, Rares Ambrus, Thomas Kollar, Zsolt Kira, and Adrien Gaidon. Shapo: Implicit representations for multi-object shape, appearance, and pose optimization, 2022. 2, 3

[26] Hamid Izadinia, Qi Shan, and Steven M. Seitz. Im2cad, 2017. 2

[27] Siyu Jiao, Hongguang Zhu, Jiannan Huang, Yao Zhao, Yunchao Wei, and Humphrey Shi. Collaborative vision-text representation optimizing for open-vocabulary segmentation, 2024. 7, 8

[28] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, 2018. 5

[29] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2, 5

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

[31] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011. 4

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

[33] Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 6

[34] Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation, 2022. 3, 8

[35] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. Mine: Towards continuous depth mpi with nerf for novel view synthesis, 2021. 2

[36] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip, 2023. 8

[37] Xiao Lin, Wenfei Yang, Yuan Gao, and Tianzhu Zhang. Instance-adaptive and geometric-aware keypoint learning for category-level 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21040–21049, 2024. 7

[38] Jian Liu, Wei Sun, Hui Yang, Jin Zheng, Zichen Geng, Hossein Rahmani, and Ajmal Mian. Monodiff9d: Monocular category-level 9d object pose estimation via diffusion model. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025. 6, 7

[39] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 6

[40] 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, 2020. 3

[41] 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, 2020. 2

[42] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 5

[43] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler, 2025. 8

[44] Stefan Popov, Pablo Bauszat, and Vittorio Ferrari. Corenet: Coherent 3d scene reconstruction from a single rgb image, 2020. 2

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2, 3

[46] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction, 2021. 5

[47] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 6

[48] 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, 2021. 2, 6

[49] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation, 2022. 6

[50] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proceedings of the 12th European Conference on Computer Vision (ECCV)*, pages 746–760, Berlin, Heidelberg, 2012. Springer. 6

[51] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 567–576, 2015. 6

[52] Stanislaw Szymanowicz, Chrisitian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10208–10217, 2024. 2, 3

[53] Stanislaw Szymanowicz, Eldar Insafutdinov, Chuanxia Zheng, Dylan Campbell, Joao F Henriques, Christian Rupprecht, and Andrea Vedaldi. Flash3d: Feed-forward generalisable 3d scene reconstruction from a single image. In *2025 International Conference on 3D Vision (3DV)*, pages 670–681. IEEE, 2025. 2

[54] Stanislaw Szymanowicz, Eldar Insafutdinov, Chuanxia Zheng, Dylan Campbell, João F. Henriques, Christian Rupprecht, and Andrea Vedaldi. Flash3d: Feed-forward generalisable 3d scene reconstruction from a single image, 2025. 2, 3

[55] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation, 2024. 2, 3

[56] Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. Diffuscene: Denoising diffusion models for generative indoor scene synthesis, 2024. 4

[57] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs, 2017. 2

[58] Shubham Tulsiani, Tinghui Zhou, Alexei A. Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency, 2017. 2

[59] Shubham Tulsiani, Richard Tucker, and Noah Snavely. Layer-structured 3d scene inference via view synthesis, 2018. 2

[60] 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, 2022. 6

[61] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991. 4

[62] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2642–2651, 2019. 2, 4

[63] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation, 2019. 3, 6

[64] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer, 2025. 3, 5, 8

[65] Yunsong Wang, Tianxin Huang, Hanlin Chen, and Gim Hee Lee. Freesplat: Generalizable 3d gaussian splatting towards free-view synthesis of indoor scenes, 2024. 2, 3

[66] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects, 2024. 6

[67] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Proceedings of Robotics: Science and Systems (RSS)*, 2018. 6

[68] Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth, 2025. 2

[69] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models, 2023. 8

[70] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation, 2023. 7, 8

[71] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation, 2024. 2, 3

[72] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2, 2024. 6, 8

[73] Kaixin Yao, Longwen Zhang, Xinhao Yan, Yan Zeng, Qixuan Zhang, Wei Yang, Lan Xu, Jiayuan Gu, and Jingyi Yu. Cast: Component-aligned 3d scene reconstruction from an rgb image, 2025. 2, 3, 4, 6

[74] Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images, 2024. 2, 3, 5

[75] Yang You, Kai Xiong, Zhening Yang, Zhengxiang Huang, Junwei Zhou, Ruoxi Shi, Zhou Fang, Adam W. Harley, Leonidas Guibas, and Cewu Lu. Pace: A large-scale dataset with pose annotations in cluttered environments, 2024. 2, 6

[76] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images, 2021. 2

[77] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip, 2023. 8

[78] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. New crfs: Neural window fully-connected crfs for monocular depth estimation, 2022. 8

[79] Cheng Zhang, Zhaopeng Cui, Yinda Zhang, Bing Zeng, Marc Pollefeys, and Shuaicheng Liu. Holistic 3d scene understanding from a single image with implicit representation, 2021. 2

[80] Chubin Zhang, Hongliang Song, Yi Wei, Yu Chen, Jiwen Lu, and Yansong Tang. Geolrm: Geometry-aware large reconstruction model for high-quality 3d gaussian generation, 2024. 2

[81] Jiyao Zhang, Weiyao Huang, Bo Peng, Mingdong Wu, Fei Hu, Zijian Chen, Bo Zhao, and Hao Dong. Omni6dpose: A benchmark and model for universal 6d object pose estimation and tracking, 2024. 2, 6

[82] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting, 2024. 2, 3

[83] Qingcheng Zhao, Xiang Zhang, Haiyang Xu, Zeyuan Chen, Jianwen Xie, Yuan Gao, and Zhuowen Tu. Depr: Depth guided single-view scene reconstruction with instance-level diffusion, 2025. 3, 4

[84] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 6

[85] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 6