

VisTa3D: A Dataset and Benchmark for Thin Object Reconstruction from Vision, Tactile, and 3D Point Clouds

Shania Guo¹, Yeongsik Seo², Andrew Fu¹, Mei Hao², Iris Xia¹,
Jiwon Jenny Lee¹, Xinyi Mary Xie¹, Hyungseob Park¹,
Aaron Dollar², and Alex Wong¹ 

¹ Yale Vision Laboratory, Yale University, New Haven, CT 06520, USA
{shania.guo, andrew.fu, jiwon.jenny.lee, x.xie, hyungseob.park, alex.wong}@yale.edu

² GRAB Lab, Yale University, New Haven, CT 06520, USA
{yeongsik.seo, mei.hao, aaron.dollar}@yale.edu

Abstract. State-of-the-art 3D reconstruction models, whether from visual, range, or both, tend to underperform on thin objects. This is partially due to the small amount of space such objects occupy in RGB images and in 3D point clouds. To test the extent of their errors, we collected the first thin object dataset comprising synchronized RGB images, depth maps, and tactile response maps, where each frame is associated with inertial measurements, camera pose and calibration, and ground-truth depth and segmentation maps obtained from laser scanning of thin objects. We hypothesize that tactile data can aid in the reconstruction of thin objects as their response maps provide local shape and deformation information. Our dataset, termed **VisTa3D**, comprises of 387 scenes covering 70 thin objects over 17 environments. We benchmarked current 3D reconstruction models on VisTa3D and found that, indeed, they exhibit low fidelity on thin objects. To test if tactile data can help, we introduce the first visual-range-tactile 3D reconstruction model as a baseline. Code and data: <https://huggingface.co/datasets/shaniaguo/VisTa3D>.

1 Introduction

Three-dimensional (3D) reconstruction of thin objects from visual and range data is challenging. The projection of thin objects onto the RGB image only occupies a small number of pixels; likewise, their returns from range sensors also tend to be sparse. This is partially due to the limited space such objects occupy in 3D as well as the finite resolution and sampling of commercial sensors, especially at longer distances. As such, even small sensor motions may cause parts of the object to be occluded (or disoccluded), which leads to a loss of correspondence across viewpoints and thus degrades reconstruction quality. Although high-resolution sensors may better capture thin objects, they are impractical in cost and require larger computational resources during downstream processing.

Recent 3D reconstruction models, trained on large-scale datasets, have generalized well across diverse 3D scenes. Despite the wide range of existing benchmark datasets, none focus on thin structures. To test the extent in which thin

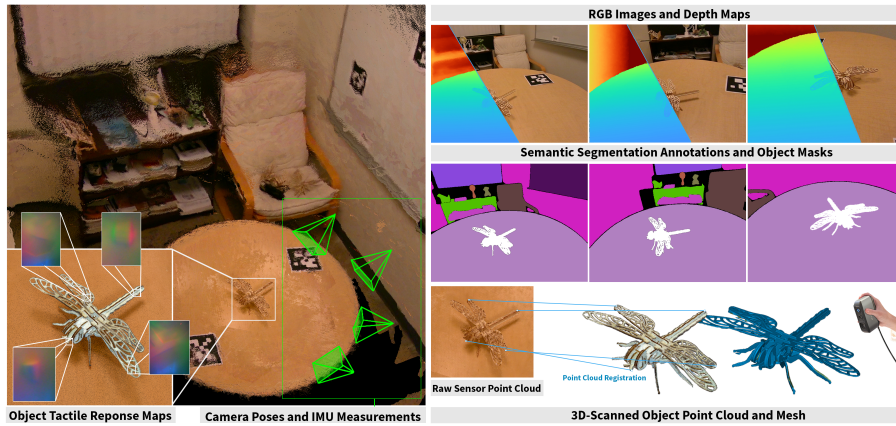


Fig. 1: *VisTa3D* comprises synchronized RGB images, depth maps, tactile response maps, and inertial measurements with camera poses, and ground truth depth maps and segmentation masks derived from 3D scanned object point clouds and meshes.

objects pose a challenge to current 3D reconstruction models, e.g., monocular depth estimation, multi-view stereo, depth completion and neural rendering methods, we collected a dataset of diverse real-world objects in both indoor and outdoor environments. We found that these models exhibit pronounced errors on thin structures. We hypothesize this is due to the limited representation of thin objects in existing training sets, as well as the limited space they occupy.

While commercial visual and range or depth sensors often fail to capture thin structures, tactile sensing offers local measurements of surface shape and deformation that are largely independent of object thickness. Tactile sensors are also complementary to visual and range sensors. At close distances, occlusion becomes more prominent and can obstruct visual observation of thin objects, while the object may also fall below the minimum sensing range of depth sensors. In contrast, tactile sensing remains effective under these conditions.

To test the potential of tactile sensing in conjunction with visual and range inputs, we introduce the first visual–range–tactile dataset for 3D reconstruction of thin objects. Our dataset includes synchronized RGB images, depth maps, and tactile response maps with inertial measurements, camera poses and calibration, and ground-truth depth maps and segmentation masks. To capture the intricate details of thin objects, we use a high-fidelity laser scanner to reconstruct each object as a 3D digital asset. We use a commercial sensor for raw data capture, which is aggregated and refined. Then, we register and insert the digital assets into the reconstruction to yield ground truth. As a byproduct, this also produces high-fidelity segmentation masks corresponding to each thin object. We organize our dataset as a benchmark aimed to test performance on thin objects. We further collect a synthetic dataset of the same data modalities to seed the (pre)training of methods for this benchmark. Fig. 1 shows an overview of the proposed dataset and benchmark of **V**ision, **T**actile, and **3D** point clouds for

thin object reconstruction, or **VisTa3D** for short. Our dataset comprises 387 scenes covering 70 thin objects over 17 environments, both real and synthetic.

As a step towards enabling high-fidelity reconstruction through the fusion of visual, tactile, and range modalities, we further propose the first (baseline) model that incorporates all three modalities. Inspired by design of depth completion networks, we found that a naive integration of tactile into visual-range fusion improves on thin object reconstruction when trained only with the synthetic portion of VisTa3D, and even more when trained on synthetic and real data.

Our contributions: (1) We introduce the first multimodal dataset for thin object reconstruction comprising synchronized RGB images, depth maps, and tactile response maps with inertial measurements, camera calibration and poses, and associated ground-truth depth maps and segmentation masks. (2) We propose an extensive data curation pipeline that involves high-fidelity laser scanning of thin objects and postprocessing that derives the above quantities. (3) We benchmark 11 existing methods on our dataset under 6 evaluation modes and found that they consistently struggle on thin objects. (4) To address this, we propose the first vision-tactile-range 3D reconstruction model as a baseline.

2 Related Work

Recent advances in 3D reconstruction, e.g., neural rendering, feed-forward models, have achieved high-fidelity outputs; yet, thin structures remain challenging.

Monocular Depth Estimation (MDE) [17, 19–24, 33, 35, 36, 69, 80, 82] aims to infer a depth map from a single image. Recent works have leveraged large-scale datasets to train foundation models [50, 51, 53, 54, 86, 87], capable of generalizing across diverse 3D scenes, and have set the state of the art. [54] demonstrated generalizable MDE through mixed-dataset training and utilized a gradient-matching loss to preserve fine details. [53] introduced a Transformer-based dense decoder. DepthAnythingV1 [86] proposed to train on pseudolabels generated by a pre-trained seed model. DepthAnythingV2 [87] extended [86] to using synthetic data to train the seed model to capture finer details. UniDepth V1 [51] introduced a pseudo-spherical representation to generalize across cameras and utilized an edge-guided loss to maintain sharp transitions; UniDepth V2 [50] simplified the architecture of [51]. Yet, little work has benchmarked MDEs on thin structures.

Multi-View Stereo (MVS) [15, 25, 58, 71, 75, 76, 88] reconstructs the 3D scene from multiple images. Traditional MVS methods employing photometric consistency struggle fundamentally with thin objects due to correspondence ambiguity [84]. [58, 76] recognize thin structures as persistent challenges where matching ambiguities arise. Learning-based methods have evolved from cost volume approaches [88] to cascade formulations enabling high-resolution reconstruction [25]. [71] proposed adaptive spatial cost aggregation for boundary stability. [96] fused feature across resolutions by multiple branches. [75] introduced the first feedforward 3D reconstruction model. Foundation models [31, 72] follow [75] and extend to large-scale datasets to generalize across diverse 3D scenes. VGGT [72] leverages global attention in transformers for feed-forward 3D infer-

ence; MapAnything [31] introduces additional modalities by tokenizing inputs. Neither have been specifically evaluated on thin objects.

Monocular Depth Completion (MDC) [18, 30, 37, 40, 42, 47, 49, 57, 77, 78, 81, 95, 99] infers a dense depth map from an image and sparse depth map. Depth completion is particularly challenging for thin objects as few points will be returned by a range sensor. CPSN [8, 9] proposed spatial propagation. NLSPN [49] extended [8] to non-local propagation while DySPN [40] to dynamic propagation. VOICED [79] and BPNet [66] prefilled sparse depth map with scaffolding and bilateral propagation. [11, 48] proposed test-time adaptation while [56, 60] performed completion on extremely sparse radar points. [30] fused image and sparse depth via a cost volume. [95] leveraged local and global context with a transformer. OMNI-DC [99] achieved zero-shot generalization across sparse depth patterns through multi-resolution depth integration with predicted depth maps from foundation monocular depth estimators as guidance.

Novel View Synthesis (NVS) [2–4, 16, 28, 32, 45, 70] reconstructs 3D scenes from multiple images. NeRF [45] introduced coordinate-based neural radiance fields optimized through differentiable volume rendering. [2] and [3] addressed aliasing with multiscale representations, while [46] accelerated optimization using multiresolution hash encodings. [4] combined grid-based representations with multi-sampling-based prefiltering to reduce z-aliasing, where thin structures may disappear. However, volumetric neural representations can still blur or miss thin structures due to lack of correspondence, as they tend to be occluded (or disoccluded) from view with small camera motion. 3D Gaussian Splatting [32] introduced explicit anisotropic Gaussian primitives for real-time rendering, but volumetric Gaussians remain prone to self-occlusion artifacts around thin structures. 2D Gaussian Splatting [28] represented scenes with oriented planar disks to improve view-consistent surface modeling, yet extremely thin structures remain challenging when optimization provides weak or inconsistent constraints.

Tactile Sensing and Reconstruction [1, 5, 12, 14, 29, 43, 61–63, 67, 74, 85, 97, 98] use contact-based local surface measurements, e.g., vision-based tactile sensors [90], for reconstruction. Prior works have used tactile sensing as a standalone modality [12, 29, 43] or combined it with complementary signals such as vision [74, 97], depth [62], and robot kinematics [5, 61]. TouchSDF [12] proposed an implicit signed distance representation. [14] unified visual and tactile observations within a shared radiance field. Gaussian-splatting approaches incorporate tactile cues into explicit scene representations: [63] supervised 3D Gaussian Splatting with fused tactile depth and uncertainty maps, while [13] injected local tactile responses into Gaussian optimization. These methods show that tactile data can provide reliable local information for neural 3D representations, but their effectiveness for thin-object reconstruction remains underexplored.

3D Reconstruction of Thin Structures remains challenging due to insufficient sampling, lack of correspondences across views, and self-occlusion (e.g., tubular objects and tree branches). [6, 10, 26, 27, 38, 41, 44, 52, 64, 68, 73, 91] study reconstruction methods for thin surfaces, curves, and tubular structures. [26] reconstructs thin surface regions, [52] provide robust mesh representations, and

[27] better preserve thin surface sections. Other approaches exploit structural priors for thin features. [6] reconstruct tubular objects through contour connections, [44] targets thin tubular structures, [38] recover thin manifold surfaces, [41] reconstruct filament-like structure from dense depth, and [73] estimates continuous 3D curves and radii. Methods to improve thin-object reconstruction include multi-view depth fusion [68], silhouette probability maps [64], densely sampled light fields [91], and signed nearest-neighbor losses for SDF learning [10].

Thin Object Datasets. Existing datasets on thin structures predominantly target 2D image segmentation. ThinObject-5K [39] provides large-scale annotations for thin object segmentation (5,743 objects), including bicycle spokes, wires, and cables. For powerline segmentation, Mendeley Powerline [89] offers paired visible-light and infrared (IR) images, while PLDU and PLDM [93] provide UAV-captured RGB images with dense segmentation labels. Beyond image segmentation, Zagar *et al.* [92] introduce point cloud datasets for thin structures, PointWire (real-world wiring harness scans) and PointVessel (synthetic tubular vessel geometry); however, they are limited to only point clouds.

In contrast, we introduce a *multimodal 3D thin object reconstruction* benchmark with visual, range, inertial and tactile data that are associated with camera poses and calibration, both real and synthetic. We provide high-fidelity ground-truth depth maps and segmentation masks derived from 3D-scanned objects. We evaluate state-of-the-art 3D reconstruction methods on our dataset and show that existing methods remain limited on thin structures.

3 Data Collection Methodology

Our data collection platform includes an Intel RealSense D455 for RGB images, depth maps, and inertial measurements, DIGIT tactile sensor for vision-based tactile response maps, and a 3DMakerPro Moose, a structured-light 3D scanner, for high-precision point clouds. Note: RGB images and depth maps share the same resolution; depth maps are registered to RGB image frame.

Vision, Inertial, Range Data Collection. We set up each scene by placing a single object on a support surface, e.g., table, bench, in a selected environment. Each sequence comprise of a single trajectory around the scene to capture a near-complete 360° view comprising both the object and environment. RGB images, (raw) depth maps, and inertial streams are recorded; inertials are synchronized with RGB images and depth maps using nearest-neighbor timestamp matching.

Given the captured RGB images and raw depth maps, we estimate camera poses using ORB-SLAM3 [7] RGB-D mode. This yields a set of camera-to-world poses $\{g_{\text{world} \leftarrow \text{cam}}^{(t)}\}_{t=1}^T$ for T frames. We manually inspect the estimated trajectories and excluded or reprocessed sequences with tracking failures. Trajectories are used in all subsequent stages, e.g., point cloud construction, object registration and insertion, and ground truth generation. Note: inertials from the Intel RealSense D455 are recorded, synchronized to RGB images and depth maps, and released, but are not used for pose estimation in our processing pipeline.



Fig. 2: 3D scans of representative thin objects from our real-world dataset.

As is common for hand-held, commodity depth sensors, the raw depth maps are noisy and are missing measurements near object boundaries. Depth maps tend to miss thin structures; (semi-)reflective surfaces and challenging lighting conditions also cause erroneous depth values. When aggregated as point clouds, raw depth maps lead to noisy, incomplete, and distorted 3D reconstructions of both the target objects and the surrounding (background) environment.

Generating Ground-truth Depth Maps. In order to obtain ground-truth 3D reconstructions of thin objects, we propose a data processing pipeline that (1) obtain a high-fidelity reconstruction (point cloud) of the thin object via a structured-light 3D scanner, (2) register the high-fidelity reconstruction to the raw noisy and incomplete point cloud of the object, (3) insert high-fidelity reconstruction into the raw point cloud (effectively replacing the noisy object) to yield a refined point cloud, and (4) re-project the refined point cloud onto the image plane corresponding to each camera pose to obtain ground-truth depth maps. Note: there are multiple filtering steps to reject outliers and clean the raw point cloud, in addition to insertion, to yield the refined point cloud.

(1) *Thin object 3D scanning.* To capture the fine structural details of the thin object, we use the 3DMakerPro Moose, a high-precision single-frame structured-light 3D scanner, to scan each object. The object is placed on a turntable for the capture. The scanner produces a high-fidelity, dense point cloud (see Fig. 1 and 2). Any missing regions are identified by manual inspection; the object is further scanned and the resulting point cloud is merged and cleaned by manual removal of outliers. We denote the final point cloud as P_{scan} .

(2) *Point cloud registration.* To register P_{scan} to the aggregated raw point cloud robustly, we first isolate the subset of points corresponding to the object of interest in the raw point cloud; we then compute the transformation between the scanned object to the object. To do so, we segment the object of interest using SAM2 [55] in each of the RGB images, which are spatially aligned with the raw depth maps. This yields a set of binary masks $\{M^{(t)}\}_{t=1}^T$, where $M^{(t)} \in \{0, 1\}^{H \times W}$ denotes the object mask for frame at time t with height H and width W . To reduce artifacts while retaining sufficient coverage of the 3D scene, we uniformly sample frames from each sequence and discard frames for which more than 50% of all depth pixels are invalid. For each depth map $z^{(t)}$ at t , we select the depth pixels within $M^{(t)}$, backproject them into 3D using the camera intrinsics

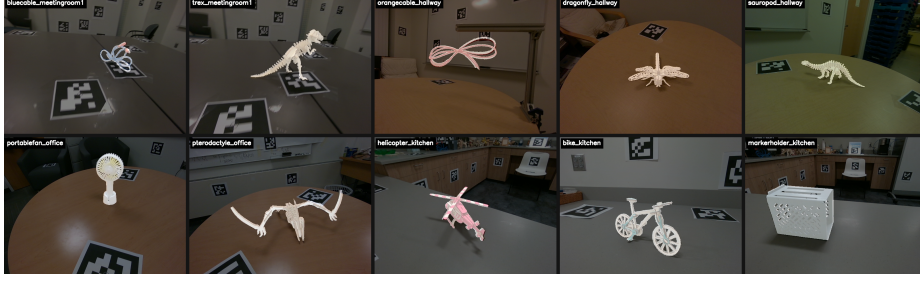


Fig. 3: *Segmentation masks.* Reprojected point clouds are overlaid as masks onto RGB images. This demonstrates the result of our highly accurate point cloud registration.

$K \in \mathbb{R}^{3 \times 3}$, and transform the resulting points into world coordinates using the estimated camera-to-world pose. Aggregating these masked points across frames yields a raw point cloud of the object, denoted P_{obj} , which serves as the registration target for aligning P_{scan} to the captured 3D scene.

$$P_{\text{obj}} = \bigcup_{t=1}^T g_{\text{world} \leftarrow \text{cam}}^{(t)} K^{-1} \bar{x} z^{(t)}(x), \quad (1)$$

where $x \in \{u \in \Omega \mid M^{(t)}(u) = 1\}$ denotes the image coordinates of the object, $\bar{x}$ their homogeneous coordinates, and Ω the image space. Since P_{obj} contains floating artifacts, we apply statistical outlier removal [59] before registration. See Supp. Mat. for more details.

Given that P_{scan} and P_{obj} are both in metric scale, we perform a two-stage rigid registration to transform P_{scan} to world coordinates via $g_{\text{world} \leftarrow \text{scan}}$. To determine correspondence between the two point clouds, we manually select three or more pairs of corresponding points between P_{scan} and P_{obj} . These sparse correspondences are used to estimate a rigid transformation, which provides an initial coarse alignment for subsequent refinement. Second, we utilize this as an initialization for Iterative Closest Point (ICP) algorithm, which iteratively establishes closest-point correspondences and updates the rigid transform to minimize point-to-point distances between P_{scan} and P_{obj} , yielding $g_{\text{world} \leftarrow \text{scan}} \in SE(3)$.

(3) *Object insertion.* Given the set of backprojected points P_{raw} from the raw depth maps, we remove the points corresponding to the object P_{obj} and insert the scanned object. This yields the final refined point cloud:

$$P = (P_{\text{raw}} \setminus P_{\text{obj}}) \cup \{g_{\text{world} \leftarrow \text{scan}} p \mid p \in P_{\text{scan}}\}. \quad (2)$$

(4) *Reprojection of refined point cloud as ground truth.* For each frame at t , we project P as depth map using the camera intrinsics K and the inverse of the camera-to-world pose, $g_{\text{cam} \leftarrow \text{world}}^{(t)}$, and rasterize it as ground truth $\hat{D}^{(t)} \in \mathbb{R}_+^{H \times W}$:

$$\hat{D}^{(t)} = \mathcal{R} \left(K g_{\text{cam} \leftarrow \text{world}}^{(t)} P \right), \quad (3)$$

where $\mathcal{R}$ denotes canonical perspective projection and rasterization. This also produces a thin object projection mask $\hat{M}^{(t)} \in \{0, 1\}^{H \times W}$, which also serves as

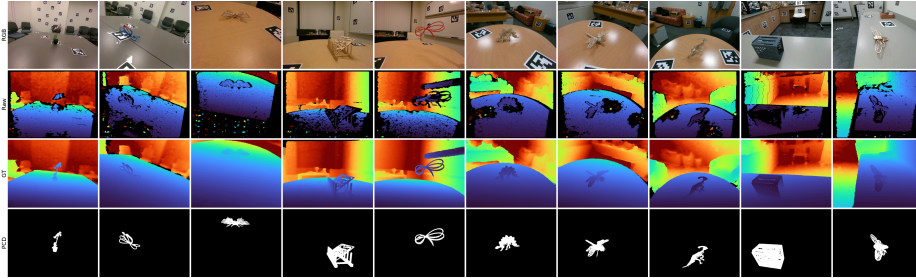


Fig. 4: *Ground-truth depth map and object segmentation masks.* Each object is scanned and inserted into an aggregated and refined point cloud to yield ground truth.

a high-quality ground-truth object segmentation (see Fig 3). Note that we use two treatments of points inside and outside the object mask: Inside $\hat{M}^{(t)}$, we apply a standard z -buffer and retain the nearest projected depth at each pixel, which preserves thin structures and prevents background points from overwriting the object; outside $\hat{M}^{(t)}$, we use a k -median buffer, where for each pixel, we collect the closest k projected depth values and assign their median value. This suppresses floating artifacts and produces smoother background surfaces and object boundaries. See Fig. 4 for representative examples.

Tactile Data Collection. We collect vision-based tactile data using a DIGIT tactile sensor. For each object, we manually select eight contact locations that cover representative local surface properties, including edges, corners, and distinctive textures. We first annotate these contact locations on the registered high-fidelity 3D scan P_{scan} , which are transformed to 3D coordinates in the world coordinate frame via Eq. 2. At each selected location, the DIGIT sensor is manually pressed against the object surface while recording a short tactile video of ≈ 200 frames. Each recording captures light- to full-forced press contact state represented as 3-channel images, and reflects the local surface deformations.

Semantic Segmentation Annotations. We annotate each sequence for semantic segmentation using CVAT, an online data annotation tool. For all scenes, we predefine a set of labels, e.g., furniture (shelf, table), room structures (wall, floor), and common indoor/outdoor objects (bench, printer, whiteboard). Annotators upload the captured RGB image sequence and manually annotate key frames using the built-in Segment Anything interactor. The masks are then propagated through the sequence using the SAM2 tracker and manually corrected for tracking failures, before being exported as semantic segmentation annotations.

4 Dataset, Benchmark, and a Baseline

VisTa3D consists of synchronized RGB images, depth maps, and tactile response maps, each associated with inertial measurements, camera poses, camera calibration, ground-truth depth maps, and segmentation masks (Fig. 1). Each high-fidelity 3D scanned object is available as point cloud and mesh forms. While VisTa3D aims to benchmark thin object 3D reconstruction, it can also support

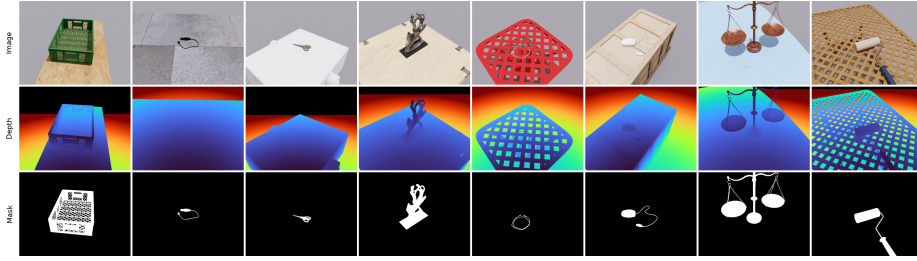


Fig. 5: *The synthetic dataset.* Examples of RGB images, depth maps, and object masks.

segmentation and localization. It consists of synthetic and real-world components with the benchmark being a subset of the real-world component; the remaining data are available for training and pretraining (e.g., on synthetic data).

Synthetic Data. The synthetic component is constructed in NVIDIA Isaac Sim 5.0. Each scene is created by arranging 18 thin objects on 1 of 9 tabletop configurations, resulting in a total of 162 scenes. For each scene, we provide RGB images, depth maps, camera intrinsics, camera poses, IMU measurements, object masks, tactile response maps from eight manually selected tactile contact locations, and object meshes (see Fig. 5). Each scene contains 500 frames covering a 360-degree view of the tabletop setup. To avoid overly regular turntable trajectories, we randomize camera elevation and distance, producing diverse viewpoints while maintaining object visibility. Objects comprise of wires, cables, figurines, and common household objects (scissors, paint roller, basket, corner shelf), and are modeled as rigid. The virtual camera was calibrated to match the Intel RealSense D455 used in the collection of real-world data.

Real-World Data. We collected 52 thin objects in 8 different environments, including 6 indoor (meeting room, hallway, office, kitchen, cubicle, lobby) and 2 outdoor (bench, staircase) locations with varying background, lighting conditions, and motion trajectories – resulting in 225 scenes in total. Our objects include wires, cables, wooden figurines (e.g., dinosaurs, insects, buildings, vehicles), plastic plants, and common office and household objects (e.g., wireless charger, portable fan, phone stand, dish rack). For each object, we provide a high-fidelity 3D scan (see Fig. 2) using 3DMakerPro Moose, a high-precision structured-light scanner with 0.03 mm scanning accuracy, 0.07 mm resolution, and 15–1500 mm working range. Each point cloud contains 5–10M valid points. Each sequence contains 300–1000 frames covering ≈ 360 degrees around the object and surrounding environment. The Intel RealSense RGB image and depth map streams are synchronized at 10 Hz with 480×640 resolution; note that the effective range of the depth values is between ≈ 0.2 m and ≈ 5 m. Inertial streams comprise of gyroscope (angular velocity) and accelerometer (linear acceleration), each represented as a vector in $\mathbb{R}^3$, recorded at 200 Hz. They are synchronized to the image frames using nearest-neighbor timestamp matching.

The Benchmark. The VisTa3D benchmark focuses on 3D reconstruction of thin objects and covers multiple reconstruction paradigms, including MDE, MVS, MDC and NVS. Only our real data component is used for benchmarking,

while the synthetic component is used for pretraining. We benchmark the baseline methods based on two tasks: depth estimation as the primary and novel view synthesis as the secondary task. For depth estimation, we split the data at the sequence level, randomly selecting 37 sequences from the full set of 225 sequences for testing and using the remaining sequences for training. This results in 88698 training frames and 13204 testing frames in total. For novel view synthesis, we instead split frames within each sequence by uniformly selecting one out of every five frames for training and using the remaining frames for testing.

Evaluation Protocol. We evaluate the predicted depth maps using standard metrics against ground truth: A1–A3 (with thresholds 1.05, 1.05², 1.05³), AbsRel, MAE, logMAE, logRMSE, and RMSE. We evaluate depth only within the range from 0.2 m to 5.0 m, corresponding to the reliable sensing range of the Intel RealSense D455. For NVS methods, we also report PSNR, SSIM, and LPIPS for rendered RGB images. See Supp. Mat. for definitions of all metrics.

As depth estimators can be scale-less or metric-scale, we evaluate each under 3 protocols to mitigate confounding factors brought by scale. *Default* reports metrics on the raw model outputs. *Median Scale* rescales outputs per frame using a median-based scale factor and *Linear Fit* applies a per-frame scale and shift – both estimated from the ground truth. To focus on thin structures, we consider 2 evaluation modes: *general* (without masking) and *thin-object focused* (masked). In the *general* evaluation, we compute errors over all pixels with valid ground truth. In thin-object focused evaluation, we compute errors only within the binary mask $\hat{M}^{(t)}$ of the projected object point cloud P_{scan} .

Summary of benchmark. We evaluate 11 representative methods spanning MDE, MVS, MDC, and NVS. We found that all methods struggle in *thin-object focused* evaluation. In response, we hypothesize that tactile data can provide the necessary local deformation information to aid thin object reconstruction. To this end, we propose the first 3D reconstruction method using visual, range, and tactile inputs as a baseline, bringing the total to 12 methods.

Tactile-DC: A Visual-Range-Tactile Baseline. To study whether localized tactile observations can aid in their thin object reconstruction, we introduce *Tactile-DC*, inspired by existing works in MDC [37, 49, 81]. To the best of our knowledge, this is the first baseline to incorporate RGB image, sparse depth map, and tactile responses for thin-object depth completion. By exploiting tactile cues and local connectivity of surfaces, Tactile-DC aims to recover fine structures in contact regions that are difficult to resolve from images, point clouds, or both.

Tactile-DC follows a conventional depth completion architecture and employs shallow RGB image and sparse depth map encoders. The RGB image and sparse depth latent feature maps are fused together by a convolutional layer. We additionally introduce a tactile encoder that encodes each tactile response map separately. As each tactile response map is associated with a 3D contact location in the world coordinate frame, we use camera intrinsics and poses to project 3D contact locations onto the image plane, following a calibrated projection [81].

Each tactile response map is encoded as feature map of size $D \times h \times w$, where D is its latent dimension and h and w are its spatial dimensions. We then

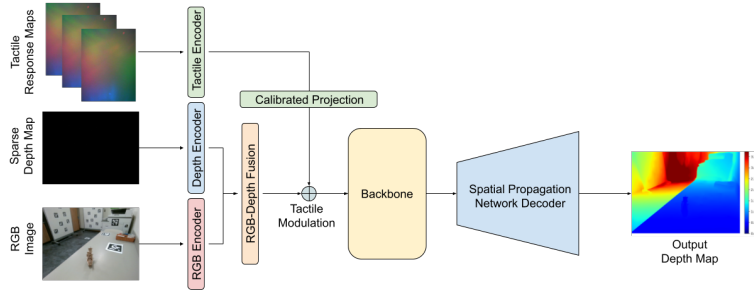


Fig. 6: *Tactile-DC*. Each input is encoded separately. RGB and sparse depth feature maps are first fused together. Tactile responses are encoded and projected onto the image plane based on contact coordinates, camera intrinsic and extrinsic calibration.

apply a 1×1 convolutional layer with batch normalization and ReLU followed by a global average pooling to compress each feature map to a $D \times 1 \times 1$ latent vector. To inject tactile information into the RGB-depth representation, each tactile latent vector is added to the locations of their projection on the image plane to modulate the fused feature maps of the RGB image and sparse depth map. Note: contact locations accounts for feature map resolution. The modulated feature map is then fed to a backbone that follows [95], and a Spatial Propagation Network [49] decodes it into a depth map. See Fig. 6 for the architecture.

5 Experiments

Implementation details, including preprocessing, model configurations, and training settings, are provided in Supp. Mat., along with ablations and analysis.

Baselines. We benchmark 12 representative methods spanning NVS (2D [28] and 3D Gaussian Splatting [32], Nerfacto [65], Depth-Nerfacto [65]), MDE (DepthAnything V1 [86], V2 [87], UniDepth V1 [51], V2 [50]), MVS (VGGT [72], MapAnything [31]), MDC (OmniDC [99]), and the proposed Tactile-DC. For all methods except Tactile-DC, we use their open-source code and checkpoints when available, and modify only data paths and input format to match our protocol.

Depth Estimation Results. Table 1 shows that baselines struggle on VisTa3D, particularly on thin-object regions. Compared with general (without masking) evaluation, thin-object focused (masked) evaluation generally leads to lower A1–A3 scores, indicating that thin object pixels are harder to reconstruct than the full image. Note: errors do not always increase under the masked setting because the thin object covers a different and often smaller depth range, but the consistent decrease in A1–A3 shows that fewer thin-object pixels are reconstructed within acceptable thresholds. Fig. 7 further illustrates this: error is higher near thin structures and object boundaries. This supports our claim that thin structures remain challenging for current reconstruction methods.

Among existing baselines, OmniDC performs best overall, especially under the default protocol. This is expected as OmniDC is a depth completion method and directly uses sparse metric depth maps as input. Qualitatively, OmniDC

Table 1: Evaluation metrics across models and protocols on the test set.

Model	Protocol	A1	A2	A3	AbsRel	logMAE	logRMSE	MAE	RMSE
General Evaluation									
UniDepth V2	Default	0.002	0.004	0.007	7.981	0.782	1.890	4.651	4.750
	Linear Fit	0.201	0.323	0.420	0.354	0.129	0.362	0.299	0.392
	Median Scale	0.152	0.265	0.359	0.395	0.164	0.477	0.420	0.582
UniDepth V1	Default	0.007	0.014	0.022	7.047	0.721	1.780	3.898	4.032
	Linear Fit	0.139	0.234	0.302	0.642	0.199	0.545	0.441	0.536
	Median Scale	0.093	0.169	0.235	0.588	0.232	0.652	0.544	0.716
DepthAnything V2	Default	0.002	0.005	0.007	4.201	0.657	1.563	3.700	4.360
	Linear Fit	0.097	0.191	0.278	0.488	0.166	0.450	0.387	0.479
	Median Scale	0.061	0.119	0.173	1.214	0.315	0.898	1.566	2.552
DepthAnything V1	Default	0.006	0.012	0.018	4.241	0.652	1.565	3.612	4.281
	Linear Fit	0.073	0.145	0.215	0.633	0.202	0.535	0.467	0.556
	Median Scale	0.054	0.106	0.156	1.093	0.303	0.856	1.314	2.189
VGGT	Default	0.060	0.117	0.172	0.486	0.215	0.545	0.496	0.633
	Linear Fit	0.229	0.409	0.550	0.202	0.086	0.279	0.175	0.261
	Median Scale	0.231	0.401	0.521	0.215	0.096	0.307	0.237	0.359
MapAnything	Default	0.511	0.627	0.676	0.307	2.096	10.965	0.210	0.495
	Linear Fit	0.242	0.376	0.452	0.393	0.137	0.447	0.254	0.395
	Median Scale	0.187	0.303	0.378	0.521	2.152	10.967	0.565	0.833
OmniDC	Default	0.834	0.881	0.897	0.409	0.054	0.312	0.139	0.460
	Linear Fit	0.484	0.569	0.624	0.367	0.102	0.320	0.212	0.323
	Median Scale	0.579	0.682	0.730	0.297	0.090	0.318	0.201	0.418
Tactile-DC	Default	0.844	0.900	0.920	0.211	0.041	0.247	0.091	0.347
	Linear Fit	0.635	0.744	0.810	0.201	0.066	0.262	0.112	0.215
	Median Scale	0.662	0.774	0.821	0.200	0.054	0.244	0.133	0.273
Object Evaluation									
UniDepth V2	Default	0.000	0.000	0.000	13.108	1.041	2.402	4.831	4.848
	Linear Fit	0.092	0.170	0.235	0.760	0.211	0.515	0.287	0.313
	Median Scale	0.080	0.162	0.243	0.695	0.195	0.466	0.273	0.290
UniDepth V1	Default	0.000	0.000	0.000	12.498	1.014	2.339	4.582	4.597
	Linear Fit	0.049	0.092	0.127	1.432	0.341	0.814	0.540	0.566
	Median Scale	0.033	0.067	0.103	1.255	0.301	0.710	0.485	0.501
DepthAnything V2	Default	0.000	0.000	0.000	4.273	0.673	1.558	1.646	1.714
	Linear Fit	0.058	0.115	0.170	0.903	0.244	0.573	0.355	0.366
	Median Scale	0.087	0.169	0.242	0.465	0.233	0.564	0.192	0.249
DepthAnything V1	Default	0.000	0.000	0.000	4.602	0.697	1.613	1.781	1.857
	Linear Fit	0.025	0.053	0.086	1.161	0.298	0.696	0.448	0.454
	Median Scale	0.071	0.140	0.205	0.554	0.237	0.574	0.227	0.281
VGGT	Default	0.064	0.125	0.186	0.563	0.205	0.484	0.214	0.228
	Linear Fit	0.119	0.235	0.360	0.321	0.104	0.297	0.141	0.230
	Median Scale	0.221	0.386	0.510	0.262	0.091	0.277	0.108	0.186
MapAnything	Default	0.165	0.288	0.375	0.485	3.307	13.607	0.195	0.293
	Linear Fit	0.035	0.072	0.113	0.668	0.207	0.520	0.262	0.302
	Median Scale	0.144	0.249	0.324	0.592	3.318	13.607	0.241	0.354
OmniDC	Default	0.687	0.772	0.817	0.274	0.058	0.260	0.101	0.238
	Linear Fit	0.342	0.423	0.472	0.513	0.142	0.390	0.194	0.263
	Median Scale	0.439	0.589	0.659	0.274	0.102	0.311	0.106	0.192
Tactile-DC	Default	0.719	0.811	0.852	0.205	0.047	0.226	0.078	0.205
	Linear Fit	0.487	0.585	0.645	0.312	0.090	0.288	0.122	0.214
	Median Scale	0.540	0.676	0.750	0.220	0.070	0.249	0.086	0.166

produces depth maps that are globally close to the ground truth, but still show considerable errors around thin structures and object boundaries. Linear fitting or median scaling does not consistently improve OmniDC as its main failure mode is not global scale mismatch but the recovery of thin structures.

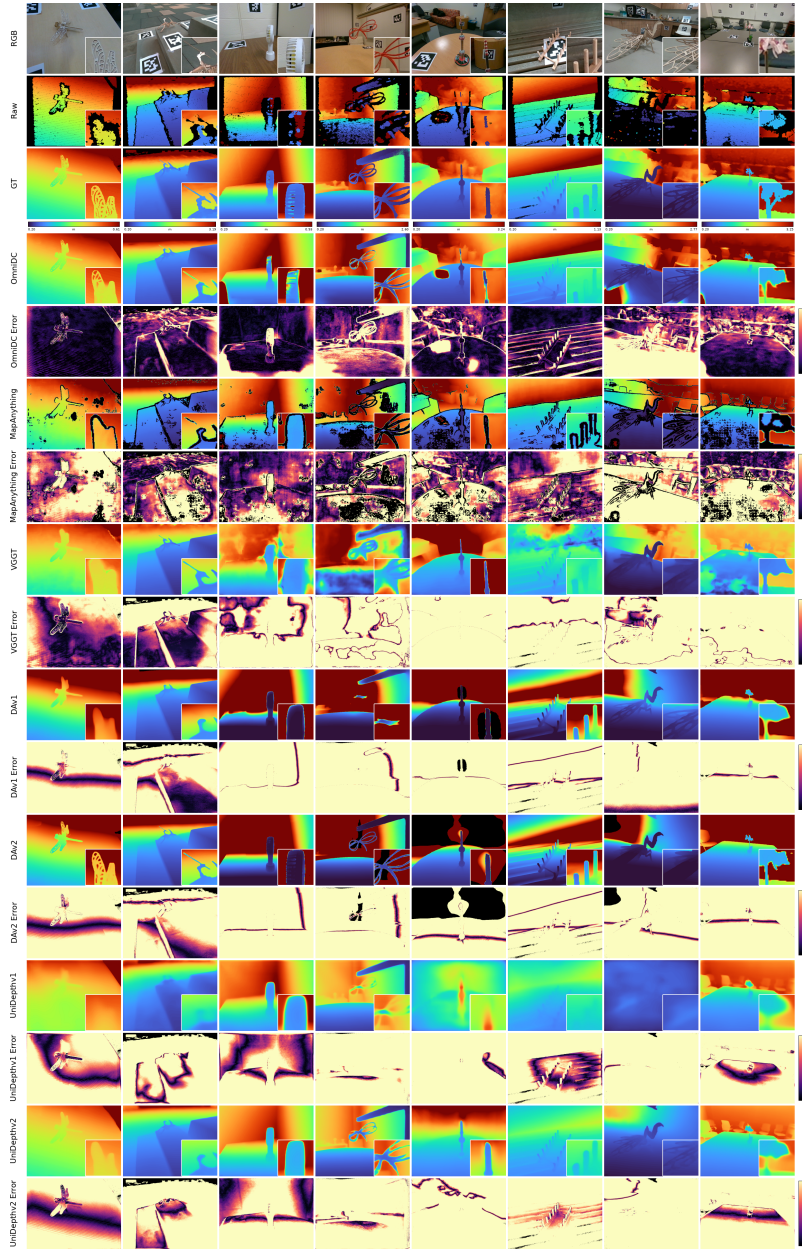


Fig. 7: Results on *VisTa3D*. Methods struggle on thin objects. Best viewed in 5 $\times$.

In contrast, MDE models such as UniDepth V1/V2, DepthAnything V1/V2 are strongly affected by scale ambiguity. All four methods perform poorly under the default protocol, especially in thin-object focused evaluation, where A1–A3 are close to zero for several settings. Linear fitting and median scaling substan-

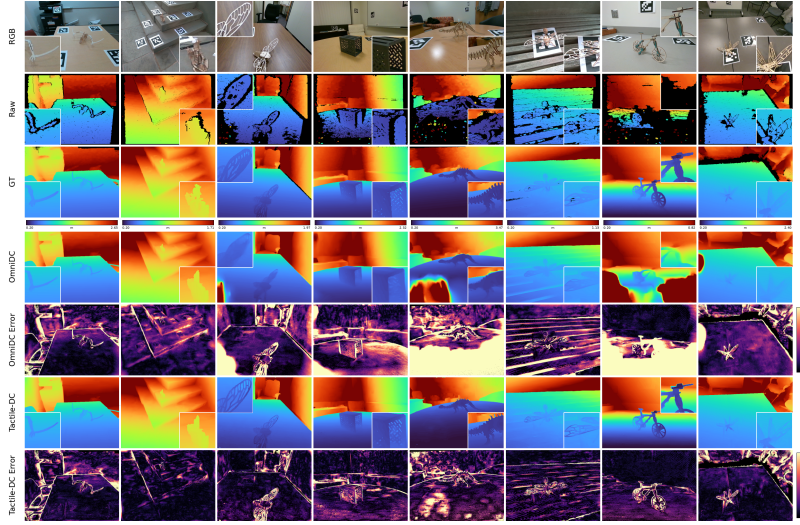


Fig. 8: Comparison between *OmniDC* and our *Tactile-DC*. *Tactile-DC* produces more accurate depth estimates, especially around thin-object areas.

tially improve their results, showing that part of the error comes from global scale mismatch. However, even after scale alignment, their accuracies and performance for the object region remain low in the qualitative results, suggesting that another error mode exists in recovering thin structures. Between model versions, UniDepth V2 generally improves over V1, and DepthAnything V2, improves over V1 after scale correction, but both still remain far below MDC (e.g., A1 below 0.10 for UniDepth V1/V2, and below 0.09 for DepthAnything V1/V2).

For MVS, VGGT performs better than MDEs, showing the benefit of multiple views; its qualitative results preserve more scene structure than MDE and exhibit smoother, more coherent outputs. However, VGGT still misses or over-smooths thin details, and its thin-object focused performance remains low. MapAnything shows a large gap between general and thin-object focused evaluation despite using images, calibration, and sparse depth, which are sufficient to ground estimates to metric scale; it also often produces noisy or incomplete estimates with scattered invalid regions and errors around thin objects and foreground-background transitions. This indicates that MVS methods may not preserve small thin-object regions reliably, even when sparse metric depth is provided.

Tactile-DC Among all methods, Tactile-DC achieves the best performance with A1/A2/A3 of 0.844/0.900/0.920 for general evaluation, and 0.719/0.811/0.852 for object evaluation. It consistently outperforms OmniDC, across all protocols under both general and thin-object focused evaluation (see Fig. 8 for comparison). In particular, Tactile-DC obtains the best A1–A3 scores and the lowest errors in all settings, demonstrating that tactile data provide useful cues for recovering thin structures that are difficult to infer from RGB images and sparse depth maps. The advantage of Tactile-DC is especially clear under thin-object focused evaluation, where visual or sparse-depth cues are often weak.

Table 2: Novel view synthesis performance on the test set.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
2D Gaussian Splatting	21.74	0.802	0.367
3D Gaussian Splatting	22.97	0.805	0.322
Nerfacto	24.16	0.855	0.177
Depth-Nerfacto	23.52	0.835	0.178

Summary: Results show a clear hierarchy amongst depth estimation methods: MDE struggle most with scale and feed-forward MVS improves global structure; both still miss thin object boundaries and details. MDC benefit from sparse metric cues, and Tactile-DC performs the best by incorporating tactile information.

NVS Results. We report PSNR, SSIM, and LPIPS in Table 2. Nerfacto achieves the best overall performance, obtaining the highest PSNR and SSIM as well as the lowest LPIPS. Depth-Nerfacto performs slightly worse despite using depth supervision, suggesting that the raw sensor depth may not consistently benefit NVS, especially when depth sensing is noisy or incomplete around thin structures. Among Gaussian-splatting (GS) methods, 3DGS outperforms 2DGS across all three metrics, indicating better rendering quality. However, GS methods are worse than Nerfacto-based methods. These results suggest that current NVS methods can synthesize plausible novel views, but sparse input views and occlusions remain challenging. Due to space limitations, we defer the NVS depth estimation results and ablations of tactile inputs to Supp. Mat.

6 Discussion and Limitations

While our dataset provides a comprehensive benchmark for thin-object reconstruction across diverse indoor and outdoor scenes, several limitations remain. First, the selection of real-world objects is constrained by the sensing characteristics of the Intel RealSense and the 3DMakerPro Moose, both of which perform unreliably on reflective, specular, or semi-transparent surfaces. Therefore, such objects are not included in our dataset. Second, despite covering a wide range of object categories, the scale of the dataset remains moderate, and real-world thin objects span a broader range of materials, thicknesses, and shapes than those represented here. Third, we assume the 3D scene is static and does not consider moving or deformable objects. Finally, MDE and MVS may benefit from task-specific finetuning, which we do not explore in this benchmark. Future work will expand object diversity, introduce controlled dynamic scenes for tasks like motion estimation [34, 83] and 4D reconstruction [94], and incorporate additional modalities to better capture failure modes of existing reconstruction pipelines.

Acknowledgements. This work is supported by NSF-2112562 Athena AI Institute and the Global Industrial Technology Cooperation Center (GITCC) through a grant agreement with the Korea Institute for Advancement of Technology (KIAT), project number P0028922.

References

1. Abad, A.C., Ranasinghe, A.: Visuotactile sensors with emphasis on gelsight sensor: A review. *IEEE Sensors Journal* **20**(14), 7628–7638 (2020)
2. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 5855–5864 (2021)
3. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 5470–5479 (2022)
4. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Zip-nerf: Anti-aliased grid-based neural radiance fields. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 19697–19705 (2023)
5. Bauza, M., Canal, O., Rodriguez, A.: Tactile mapping and localization from high-resolution tactile imprints. In: *2019 International Conference on Robotics and Automation (ICRA)*. pp. 3811–3817. IEEE (2019)
6. Bornik, A.: Reconstruction and representation of tubular structures using simplex meshes (2005)
7. Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M., Tardós, J.D.: Orb-slam3: An accurate open-source library for visual, visual-inertial, and multimap slam. *IEEE transactions on robotics* **37**(6), 1874–1890 (2021)
8. Cheng, X., Wang, P., Guan, C., Yang, R.: Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 34, pp. 10615–10622 (2020)
9. Cheng, X., Wang, P., Yang, R.: Learning depth with convolutional spatial propagation network. *IEEE transactions on pattern analysis and machine intelligence* **42**(10), 2361–2379 (2019)
10. Chou, G., Chugunov, I., Heide, F.: Gensdf: Two-stage learning of generalizable signed distance functions. *Advances in Neural Information Processing Systems* **35**, 24905–24919 (2022)
11. Chung, Y., Park, H., Rim, P., Zhang, X., He, J., Zeng, Z., Cicek, S., Hong, B.W., Duncan, J.S., Wong, A.: Eta: Energy-based test-time adaptation for depth completion. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2025)
12. Comi, M., Lin, Y., Church, A., Tonioni, A., Aitchison, L., Lepora, N.F.: Touchsdf: A deepsdf approach for 3d shape reconstruction using vision-based tactile sensing. *IEEE Robotics and Automation Letters* **9**(6), 5719–5726 (2024)
13. Comi, M., Tonioni, A., Tremblay, J., Yang, M., Blukis, V., Lin, Y., Lepora, N.F., Aitchison, L.: Snap-it, tap-it, splat-it: Tactile-informed 3d gaussian splatting for reconstructing challenging surfaces. In: *2025 International Conference on 3D Vision (3DV)*. pp. 1134–1143. IEEE (2025)
14. Dou, Y., Yang, F., Liu, Y., Loquercio, A., Owens, A.: Tactile-augmented radiance fields. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 26529–26539 (2024)
15. Duan, R., Hong, E., Turner, E., Wong, A., Zhou, Y.: Fisheye3r: Adapting unified 3d feed-forward foundation models to fisheye lenses. In: *European Conference on Computer Vision*. Springer (2026)

16. Duan, R., Wong, A.: Evidential neural radiance fields. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2026)
17. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE international conference on computer vision. pp. 2650–2658 (2015)
18. Ezhov, V., Park, H., Zhang, Z., Upadhyay, R., Zhang, H., Chandrappa, C.C., Kadambi, A., Ba, Y., Dorsey, J., Wong, A.: All-day depth completion. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE (2024)
19. Fei, X., Wong, A., Soatto, S.: Geo-supervised visual depth prediction. IEEE Robotics and Automation Letters **4**(2), 1661–1668 (2019)
20. Gangopadhyay, S., Kim, J.H., Chen, X., Rim, P., Park, H., Wong, A.: Extending foundational monocular depth estimators to fisheye cameras with calibration tokens. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025)
21. Gangopadhyay, S., Wong, A.: From perspective to fisheye depth estimation and open-vocabulary segmentation. In: European Conference on Computer Vision. Springer (2026)
22. Garg, R., Bg, V.K., Carneiro, G., Reid, I.: Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: European conference on computer vision. pp. 740–756. Springer (2016)
23. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 270–279 (2017)
24. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3828–3838 (2019)
25. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2495–2504 (2020)
26. Hilton, A., Stoddart, A., Illingworth, J., Windeatt, T.: Reconstruction of 3d delaunay surface models of complex objects. In: 1996 IEEE International Conference on Systems, Man and Cybernetics. Information Intelligence and Systems (Cat. No. 96CH35929). vol. 4, pp. 2445–2450. IEEE (1996)
27. Hilton, A., Stoddart, A.J., Illingworth, J., Windeatt, T.: Implicit surface-based geometric fusion. Computer Vision and Image Understanding **69**(3), 273–291 (1998)
28. Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2d gaussian splatting for geometrically accurate radiance fields. In: ACM SIGGRAPH 2024 conference papers. pp. 1–11 (2024)
29. Huang, H.J., Mirzaee, M.A., Kaess, M., Yuan, W.: Gelslam: A real-time, high-fidelity, and robust 3d tactile slam system. arXiv preprint arXiv:2508.15990 (2025)
30. Kam, J., Kim, J., Kim, S., Park, J., Lee, S.: Costdcnet: Cost volume based depth completion for a single rgb-d image. In: European Conference on Computer Vision. pp. 257–274. Springer (2022)
31. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction; map-anything. github. io. In: 2026 International Conference on 3D Vision (3DV). pp. 499–509. IEEE (2026)
32. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)

33. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth international conference on 3D vision (3DV). pp. 239–248. IEEE (2016)
34. Lao, D., Sundaramoorthi, G.: Extending layered models to 3d motion. In: Proceedings of the European conference on computer vision (ECCV). pp. 435–451 (2018)
35. Lao, D., Wu, Y., Liu, T.Y., Wong, A., Soatto, S.: Sub-token vit embedding via stochastic resonance transformers. In: International Conference on Machine Learning. PMLR (2024)
36. Lao, D., Yang, F., Wang, D., Park, H., Lu, S., Wong, A., Soatto, S.: On the viability of monocular depth pre-training for semantic segmentation. In: European Conference on Computer Vision. Springer (2024)
37. Li, A., Yuan, Z., Ling, Y., Chi, W., Zhang, C., et al.: A multi-scale guided cascade hourglass network for depth completion. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 32–40 (2020)
38. Li, S., Yao, Y., Fang, T., Quan, L.: Reconstructing thin structures of manifold surfaces by integrating spatial curves. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2887–2896 (2018)
39. Liew, J.H., Cohen, S., Price, B., Mai, L., Feng, J.: Deep interactive thin object selection. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 305–314 (2021)
40. Lin, Y., Cheng, T., Zhong, Q., Zhou, W., Yang, H.: Dynamic spatial propagation network for depth completion. In: Proceedings of the aaai conference on artificial intelligence. vol. 36, pp. 1638–1646 (2022)
41. Liu, L., Chen, N., Ceylan, D., Theobalt, C., Wang, W., Mitra, N.J.: Curve-fusion: Reconstructing thin structures from rgbd sequences. arXiv preprint arXiv:2107.05284 (2021)
42. Liu, T.Y., Agrawal, P., Chen, A., Hong, B.W., Wong, A.: Monitored distillation for positive congruent depth completion. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II. pp. 35–53. Springer (2022)
43. Lu, J., Wan, Z., Zhang, Y.: Tac2structure: Object surface reconstruction only through multi times touch. IEEE Robotics and Automation Letters **8**(3), 1391–1398 (2023)
44. Martin, T., Montes, J., Bazin, J.C., Popa, T.: Topology-aware reconstruction of thin tubular structures. In: SIGGRAPH Asia 2014 Technical Briefs, pp. 1–4. ACM (2014)
45. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
46. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM transactions on graphics (TOG) **41**(4), 1–15 (2022)
47. Park, H., Chen, R., Rim, P., Lao, D., Wong, A.: Orcas: Unsupervised depth completion via occluded region completion as supervision. The Fourteenth International Conference on Learning Representations (2026)
48. Park, H., Gupta, A., Wong, A.: Test-time adaptation for depth completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20519–20529 (2024)

49. Park, J., Joo, K., Hu, Z., Liu, C.K., So Kweon, I.: Non-local spatial propagation network for depth completion. In: European conference on computer vision. pp. 120–136. Springer (2020)
50. Piccinelli, L., Sakaridis, C., Yang, Y.H., Segu, M., Li, S., Abbeloos, W., Van Gool, L.: Unidepthv2: Universal monocular metric depth estimation made simpler. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025)
51. Piccinelli, L., Yang, Y.H., Sakaridis, C., Segu, M., Li, S., Van Gool, L., Yu, F.: Unidepth: Universal monocular metric depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10106–10116 (2024)
52. Pulli, K., Duchamp, T., Hoppe, H., McDonald, J., Shapiro, L., Stuetzle, W.: Robust meshes from multiple range maps. In: Proceedings. International Conference on Recent Advances in 3-D Digital Imaging and Modeling (Cat. No. 97TB100134). pp. 205–211. IEEE (1997)
53. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)
54. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence* **44**(3), 1623–1637 (2020)
55. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
56. Rim, P., Park, H., Ezhov, V., Moon, J., Wong, A.: Radar-guided polynomial fitting for metric depth estimation. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2026)
57. Rim, P., Park, H., Zeng, Z., Chung, Y., Wong, A.: Protodepth: Unsupervised continual depth completion with prototypes. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6304–6316 (2025)
58. Romanoni, A., Matteucci, M.: Tapa-mvs: Textureless-aware patchmatch multi-view stereo. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10413–10422 (2019)
59. Rusu, R.B., Cousins, S.: 3d is here: Point cloud library (pcl). In: 2011 IEEE international conference on robotics and automation. pp. 1–4. IEEE (2011)
60. Singh, A.D., Ba, Y., Sarker, A., Zhang, H., Kadambi, A., Soatto, S., Srivastava, M., Wong, A.: Depth estimation from camera image and mmwave radar point cloud. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9275–9285 (2023)
61. Suresh, S., Qi, H., Wu, T., Fan, T., Pineda, L., Lambeta, M., Malik, J., Kalakrishnan, M., Calandra, R., Kaess, M., et al.: Neuralfeels with neural fields: Visuotactile perception for in-hand manipulation. *Science Robotics* **9**(96), eadl0628 (2024)
62. Suresh, S., Si, Z., Mangelson, J.G., Yuan, W., Kaess, M.: Shapemap 3-d: Efficient shape mapping through dense touch and vision. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 7073–7080. IEEE (2022)
63. Swann, A., Strong, M., Do, W.K., Camps, G.S., Schwager, M., Kennedy, M.: Touchgs: Visual-tactile supervised 3d gaussian splatting. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 10511–10518. IEEE (2024)

64. Tabb, A.: Shape from silhouette probability maps: reconstruction of thin objects in the presence of silhouette extraction and calibration error. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 161–168 (2013)
65. Tancik, M., Weber, E., Ng, E., Li, R., Yi, B., Wang, T., Kristoffersen, A., Austin, J., Salahi, K., Ahuja, A., et al.: Nerfstudio: A modular framework for neural radiance field development. In: *ACM SIGGRAPH 2023 conference proceedings*. pp. 1–12 (2023)
66. Tang, J., Tian, F.P., An, B., Li, J., Tan, P.: Bilateral propagation network for depth completion. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9763–9772 (2024)
67. Tu, J., Yang, F., Ma, C., Yu, X., Zeng, Z., Wu, S., Zhao, H., Tao, Z., Zhang, C., Qian, H., Wong, A.: Unitac: A unified multimodal model for cross-sensor tactile understanding and generation. In: *European Conference on Computer Vision*. Springer (2026)
68. Ummenhofer, B., Brox, T.: Point-based 3d reconstruction of thin objects. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 969–976 (2013)
69. Upadhyay, R., Zhang, H., Ba, Y., Yang, E., Gella, B., Jiang, S., Wong, A., Kadambi, A.: Enhancing diffusion models with 3d perspective geometry constraints. *ACM Transactions on Graphics (TOG)* **42**(6), 1–15 (2023)
70. Wang, D., Rim, P., Tian, T., Lao, D., Wong, A., Sundaramoorthi, G.: Ode-gs: Latent odes for dynamic scene extrapolation with 3d gaussian splatting. *The Fourteenth International Conference on Learning Representations* (2026)
71. Wang, F., Galliani, S., Vogel, C., Speciale, P., Pollefeys, M.: Patchmatchnet: Learned multi-view patchmatch stereo. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 14194–14203 (2021)
72. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2025)
73. Wang, P., Liu, L., Chen, N., Chu, H.K., Theobalt, C., Wang, W.: Vid2curve: simultaneous camera motion estimation and thin structure reconstruction from an rgb video. *ACM Transactions on Graphics (TOG)* **39**(4), 132–1 (2020)
74. Wang, S., Wu, J., Sun, X., Yuan, W., Freeman, W.T., Tenenbaum, J.B., Adelson, E.H.: 3d shape perception from monocular vision, touch, and shape priors. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 1606–1613. IEEE (2018)
75. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 20697–20709 (2024)
76. Wang, Y., Huang, K., Chen, Z., Chen, W., et al.: Multi-view stereo for weakly textured indoor 3d reconstruction. *Computer-Aided Civil and Infrastructure Engineering* (2024)
77. Wong, A., Cicek, S., Soatto, S.: Learning topology from synthetic data for unsupervised depth completion. *IEEE Robotics and Automation Letters* **6**(2), 1495–1502 (2021)
78. Wong, A., Fei, X., Hong, B.W., Soatto, S.: An adaptive framework for learning unsupervised depth completion. *IEEE Robotics and Automation Letters* **6**(2), 3120–3127 (2021)
79. Wong, A., Fei, X., Tsuei, S., Soatto, S.: Unsupervised depth completion from visual inertial odometry. *IEEE Robotics and Automation Letters* **5**(2), 1899–1906 (2020)

80. Wong, A., Soatto, S.: Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5644–5653 (2019)
81. Wong, A., Soatto, S.: Unsupervised depth completion with calibrated backprojection layers. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 12747–12756 (2021)
82. Wu, Y., Liu, T.Y., Park, H., Soatto, S., Lao, D., Wong, A.: Augundo: Scaling up augmentations for monocular depth completion and estimation. In: *European Conference on Computer Vision*. Springer (2024)
83. Xiao, Y., Rodriguez Coronel, C., Zhan, J., Ehsani Oskouie, H., Wong, A., Lao, D.: Triangular consistency as a universal constraint for learning optical flow. In: *European Conference on Computer Vision*. Springer (2026)
84. Xu, Z., Liu, Y., Shi, X., Wang, Y., Zheng, Y.: Marmvs: Matching ambiguity reduced multiple view stereo for efficient large scale scene reconstruction. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 5981–5990 (2020)
85. Yang, F., Feng, C., Chen, Z., Park, H., Wang, D., Dou, Y., Zeng, Z., Chen, X., Gangopadhyay, R., Owens, A., Wong, A.: Binding touch to everything: Learning unified multimodal tactile representations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 26340–26353 (2024)
86. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 10371–10381 (2024)
87. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. *Advances in Neural Information Processing Systems* **37**, 21875–21911 (2024)
88. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 767–783 (2018)
89. Yetgin, Ö.E., Gerek, Ö.N.: Powerline image dataset (infrared-ir and visible light-vl). *Mendeley Data* **7**, 2017 (2017)
90. Yuan, W., Dong, S., Adelson, E.H.: Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors* **17**(12), 2762 (2017)
91. Yucer, K., Kim, C., Sorkine-Hornung, A., Sorkine-Hornung, O.: Depth from gradients in dense light fields for object reconstruction. In: *2016 fourth international conference on 3D vision (3DV)*. pp. 249–257. IEEE (2016)
92. Žagar, B.L., Liu, M., Hertel, T., Yurtsever, E., Knoll, A.C.: 3d understanding of deformable linear objects: Datasets and transferability benchmark. In: *Proceedings of the Winter Conference on Applications of Computer Vision*. pp. 6404–6414 (2025)
93. Zhang, H., Yang, W., Yu, H., Zhang, H., Xia, G.S.: Detecting power lines in uav images with convolutional features and structured constraints. *Remote Sensing* **11**(11), 1342 (2019)
94. Zhang, J., Herrmann, C., Hur, J., Jampani, V., Cole, F., Sun, D., Yang, M.H., et al.: Monst3r: A simple approach for estimating geometry in the presence of motion. In: *International Conference on Learning Representations*. vol. 2025, pp. 82863–82886 (2025)
95. Zhang, Y., Guo, X., Poggi, M., Zhu, Z., Huang, G., Mattoccia, S.: Completion-former: Depth completion with convolutions and vision transformers. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 18527–18536 (2023)

- 96. Zhang, Z., Peng, R., Hu, Y., Wang, R.: Geomvsnet: Learning multi-view stereo with geometry perception. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21508–21518 (2023)
- 97. Zhao, J., Bauza, M., Adelson, E.H.: Fingerslam: Closed-loop unknown object localization and reconstruction from visuo-tactile feedback. arXiv preprint arXiv:2303.07997 (2023)
- 98. Zhu, Y., Hao, M., Zhu, X., Bateux, Q., Wong, A., Dollar, A.M.: Forces for free: Vision-based contact force estimation with a compliant hand. *Science Robotics* **10**(103), eadq5046 (2025)
- 99. Zuo, Y., Yang, W., Ma, Z., Deng, J.: Omni-dc: Highly robust depth completion with multiresolution depth integration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9287–9297 (2025)