TouchAnything: Diffusion-Guided 3D Reconstruction from Sparse Robot Touches

Langzhe Gu[1,3], Hung-Jui Huang[2*], Mohamad Qadri[2*], Michael Kaess[2], and Wenzhen Yuan[3]

[1] Tsinghua University
[2] Carnegie Mellon University
[3] University of Illinois Urbana-Champaign
[*] Equal contribution

Abstract. Accurate object geometry estimation is essential for many downstream tasks, including robotic manipulation and physical interaction. Although vision is the dominant modality for shape perception, it becomes unreliable under occlusions or challenging lighting conditions. In such scenarios, tactile sensing provides direct geometric information through physical contact. However, reconstructing global 3D geometry from sparse local touches alone is fundamentally underconstrained. We present TouchAnything, a framework that leverages a pretrained large-scale 2D vision diffusion model as a semantic and geometric prior for 3D reconstruction from sparse tactile measurements. Unlike prior work that trains category-specific reconstruction networks or learns diffusion models directly from tactile data, we transfer the geometric knowledge encoded in pretrained visual diffusion models to the tactile domain. Given sparse contact constraints and a coarse class-level description of the object, we formulate reconstruction as an optimization problem that enforces tactile consistency while guiding solutions toward shapes consistent with the diffusion prior. Our method reconstructs accurate geometries from only a few touches, outperforms existing baselines, and enables open-world 3D reconstruction of previously unseen object instances.

Keywords: Tactile Sensing · 3D Reconstruction · Generative Priors

1 Introduction

Tactile feedback is a fundamental sensory signal that enables humans to interact effectively with the physical world. Touch provides rich information about the geometry and rigidity of objects, from which the nature of possible interactions (e.g., grasps, handling) can be inferred. This becomes especially important under heavy occlusions or challenging lighting conditions, where visual perception may fail. In such situations, the ability to rely on touch alone becomes essential for estimating object geometry and enabling downstream manipulation tasks.

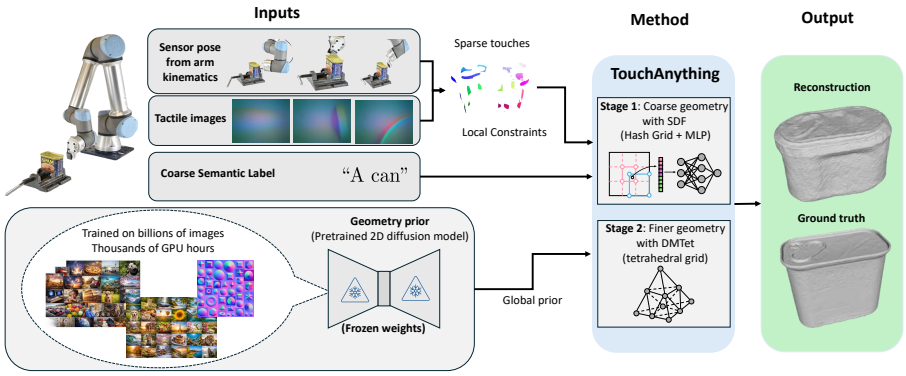

Fig. 1: Overview of TouchAnything. Sparse tactile measurements and a class-level text description are combined with a pretrained 2D diffusion model serving as a geometric prior. Local tactile constraints and global diffusion geometric guidance jointly optimize a two-stage geometric representation.

However, geometry estimation from touch alone is inherently underconstrained due to the sparsity of contact measurements. What makes this possible in humans is the presence of strong prior knowledge about object shape and structure. For example, imagine that you are tasked with finding a pen inside a backpack without visual feedback. You already possess a semantic understanding of what a "pen" typically looks like (e.g. its approximate shape and structure). As you make contact with objects inside the bag, sparse tactile cues are combined with this prior knowledge to refine your belief about the object being touched. Touch does not operate in isolation; rather, it conditions and disambiguates an existing internal model of object geometry.

This raises a natural question: can we equip robots with a similar capability? Specifically, given a robot arm equipped with tactile sensing, can we infer object geometry from sparse touches when guided by a coarse semantic prior and can this be achieved in an open-world setting? [4]

Recent work has demonstrated that diffusion models can serve as powerful priors for geometric reasoning. Diffusion-based approaches have been successfully applied to 3D reconstruction, text-to-3D generation, and multi-view consistent shape synthesis, enabling plausible 3D geometry inference even under limited observations or severely constrained settings. Importantly, these models are trained on large-scale visual datasets and have not been specialized to tactile signals. Nevertheless, they implicitly encode strong geometric information that may transfer across sensing modalities. To the best of our knowledge, this work is the first to adapt a general-purpose off-the-shelf pretrained 2D vision diffusion model as a geometric prior for 3D reconstruction from sparse tactile touches.

[4] By open-world, we mean reconstruction of previously unseen object instances without object-specific training.

In contrast to prior touch-based reconstruction methods, which train models on class-specific datasets spanning just a handful of object categories [8, 46, 52], we develop a system that leverages the geometric priors encoded in large-scale visual diffusion models trained on billions of internet images to guide 3D reconstruction given sparse tactile measurements. We argue that transferring large-scale visual priors to the tactile domain represents an important step toward open-world generalization, especially in regimes where tactile data is scarce and training specialized generative models is impractical.

Our contributions are as follows:

- We introduce **TouchAnything**, a framework for reconstructing global 3D object geometry from sparse tactile contacts and a coarse semantic prior, enabling open-world 3D reconstruction inference from limited physical interaction.
- We demonstrate that large-scale pretrained 2D vision diffusion models can be repurposed as geometric priors for tactile reconstruction, transferring visual generative knowledge to the tactile domain without task-specific diffusion training.
- We provide extensive validation in simulation and real-world robotic experiments, including a study of reconstruction accuracy under varying numbers of touches and prompt designs.

2 Related Work

2.1 3D Reconstruction from Sparse Observations

Neural implicit and differentiable geometric representations such as neural radiance fields (NeRF) [21], signed distance fields (SDFs) [25], hybrid mesh-based approaches like DMTet [38], and more recently 3D Gaussian Splatting [15], have become dominant frameworks for 3D reconstruction. These methods introduce differentiable parameterizations of geometry and appearance, making them well-suited for end-to-end optimization. Beyond RGB-based reconstruction, their flexibility has enabled applications across diverse sensing modalities [16, 20, 27, 30, 32, 33, 53], including tactile sensing [8, 41]. However, regardless of the underlying representation or sensing modality, 3D reconstruction becomes severely underconstrained when observations provide only limited geometric coverage of the underlying surface [24, 44]. In few-view settings, implicit and differentiable models alone are insufficient to solve for global geometry without additional regularization. This challenge is further amplified in tactile reconstruction, where measurements come from small local contact patches rather than partial global image observations.

2.2 Diffusion Models as Geometric Priors

Recent advances in diffusion modeling have demonstrated that large-scale generative models implicitly encode rich knowledge about object geometry. DreamFusion [26] introduced the use of pretrained 2D text-to-image diffusion models to

optimize 3D representations via score distillation sampling (SDS), enabling text-to-3D generation without direct 3D supervision. Subsequent works [19, 31, 47], improved fidelity and enhanced geometric detail in diffusion-guided 3D synthesis. Fantasia3D [6] and RichDreamer [31] explicitly incorporate geometric signals such as surface normals and depth to better disentangle shape and appearance during diffusion-guided 3D generation. These works demonstrate that diffusion models can effectively leverage explicit geometric cues to improve the quality of the generated geometry. In our setting, we similarly render normal maps from the evolving 3D geometry. However, unlike prior methods, we additionally enforce consistency with real local surface normal measurements obtained from image-based tactile sensors. We therefore combine global diffusion guidance with physical local constraints imposed by robot contact. Tactile DreamFusion [11] similarly incorporates tactile signals into text-to-3D synthesis, but its primary objective remains asset generation, whereas our goal is tactile-conditioned 3D reconstruction constrained by real tactile measurements.

Beyond text-to-3D synthesis, sparse 3D reconstruction has been addressed by learning task-specific shape completion priors from curated 3D datasets [7, 49]. More recently, diffusion-based completion models [14,37] train 3D diffusion priors to regularize underconstrained geometric inference and recover plausible structure from partial observations. Unlike these approaches, which require supervised training on 3D shape collections, we leverage a pretrained 2D diffusion model as a transferable geometric prior without task-specific diffusion training.

Several recent works [23, 48, 55] also employ diffusion models as geometric priors for sparse-view 3D reconstruction from RGB images, where diffusion guidance regularizes geometry estimated from limited visual coverage. In contrast, our setting relies solely on sparse local tactile contacts, which provide highly localized surface measurements without global image-level constraints. Inferring global geometry from such physical interaction signals is a different and a more severely underconstrained problem.

2.3 Tactile-based Shape Reconstruction

Image-based tactile sensors provide useful geometric information (e.g. contact location and surface normals) which have been leveraged for object state estimation and tracking during interaction [12, 13, 28, 29] as well as 3D reconstruction. TouchSDF [8] predicts local surface geometry from vision-based tactile sensors and learns an SDF encoding the object shape. More recently, diffusion-based tactile reconstruction methods [46, 52] train conditional diffusion models for tactile-conditioned 3D reconstruction using object-level datasets such as ShapeNet and ABC. These approaches rely on supervised training with ground-truth 3D geometry from fixed object categories (e.g., guitars, bottles), which may limit generalization beyond seen categories and involve computationally intensive task-specific training of diffusion models. Other works [9, 10, 41–43] fuse tactile and visual measurements for 3D reconstruction benefiting from the global visual coverage provided by vision. However, they differ from our setting where

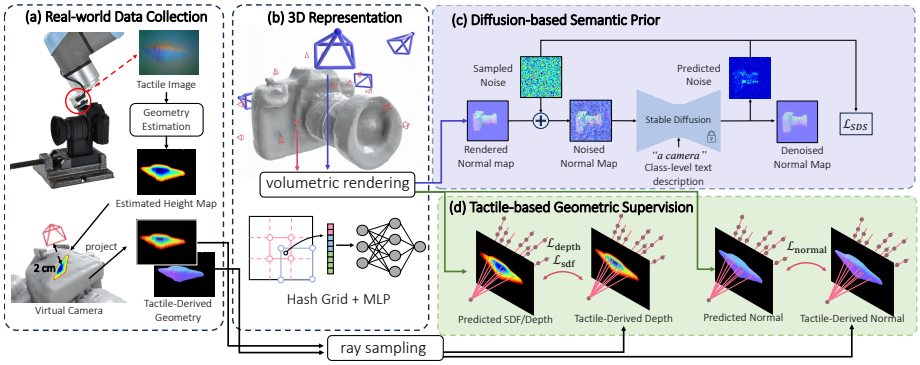

Fig. 2: TouchAnything reconstruction pipeline. We use a GelSight tactile sensor mounted on a robotic arm to collect raw tactile measurements, which are converted into local depth and surface normal maps (Sec. 3.1). These measurements provide sparse geometric supervision for learning a neural signed distance field (SDF) by enforcing local constraints through depth and normal losses. Normal maps rendered from the evolving 3D geometry are fed into a pretrained Stable Diffusion model, which provides global geometric guidance through score distillation sampling (SDS). By jointly enforcing tactile consistency combined with a pretrained 2D diffusion model and a class-level text description, the system reconstructs globally consistent 3D geometry from sparse contact measurements (Sec. 3.2).

touch provides the primary geometric signal. Overall, prior tactile reconstruction methods either depend on dense measurements, multi-modal sensing, or expensive training of task-specific generative priors. To the best of our knowledge, reconstructing global object geometry from sparse tactile contacts using pretrained off-the-shelf 2D diffusion models remains unexplored.

3 Methodology

Our goal is to reconstruct the 3D geometry of an arbitrary object using only sparse tactile readings and a minimal class-level text description (e.g. "a camera", "a bottle"). We assume that the contact locations of tactile readings are known from robot kinematics. The text description provides only a weak semantic prior at the category level, reflecting realistic scenarios where a robot knows the object class but not the specific geometry, such as searching for a key in the wallet.

To solve this problem, we propose TouchAnything, a diffusion-guided method for 3D reconstruction from sparse tactile inputs. Instead of utilizing a model trained on limited object categories [8, 52], we leverage Stable Diffusion [35], a general-purpose generative model, to guide reconstruction, conditioned on the weak text description. This design makes our method much more general, enabling reconstruction of arbitrary objects without pre-training or class-specific data collection. In TouchAnything, tactile readings are first converted into local depth maps using established methods [42, 45], and represent it as virtual

camera observations to ensure compatibility with the vision-centric reconstruction pipeline. We then adopt a coarse-to-fine reconstruction strategy inspired by Magic3D [19] and optimize the object geometry by jointly enforcing consistency with tactile-derived geometry and alignment with the class-level text description.

3.1 Deriving Local Geometry from Tactile Sensing

We use a GelSight sensor [50] to collect tactile data. GelSight is a vision-based tactile sensor composed of a soft elastomer sensing surface, an integrated illumination system, and a camera. Upon contact, the elastomer deforms, changing the reflected illumination pattern, which is captured by the camera and used to reconstruct the contact surface geometry via photometric stereo. A sample tactile image is shown in Fig. 2a. In our work, GelSight images are obtained from both real-world experiments and simulations. From these images, we estimate the local contact geometry and convert it into virtual camera observations for compatibility with the vision-centric reconstruction pipeline.

Geometry Estimation from Real-World Tactile Data Given a tactile image $\mathbf{T}$ collected by a physical GelSight sensor, we adopt the learning-based method in [45] to estimate per-pixel surface gradients from the RGB values and coordinates of each pixel using a three-layer MLP. The predicted gradients are then integrated using a fast Poisson solver [50] to recover the surface depth map. The contact mask is obtained by thresholding the estimated depth map.

Geometry Estimation from Simulated Tactile Data Given a photorealistic simulated GelSight image, we adopt the learning-based method in [42], using a multi-head U-Net [36] to jointly predict the depth map and contact mask. The network is trained on 20k simulated GelSight images generated from contacts with 78 YCB objects [4]. Compared to the real-world scenario, we use a more complex geometry estimation model for simulated tactile data because larger-scale training data are available in simulation.

Tactile-Derived Geometry as Virtual Camera Observation For each tactile reading $\mathbf{T}_i$, we place a virtual camera $\mathbf{C}_i$ 2.0 cm behind the contact patch and convert the tactile-estimated depth map and contact mask into virtual camera observations consisting of depth and normal maps with an associated contact mask. The reconstruction process then operates on these tactile-derived geometries in the form of virtual camera observations rather than the raw tactile readings, ensuring compatibility with the vision-centric reconstruction pipeline. The full data collection pipeline is shown in Fig. 2a.

3.2 Stage 1: Learning Coarse Geometry of the Object

We model the coarse object geometry using a SDF-based neural implicit representation $f_\theta : \mathbb{R}^3 \to \mathbb{R}$, implemented using a Neuralangelo-style [18] multi-resolution 3D hash-grid features followed by an MLP (Fig. 2b). The function f_θ

maps a 3D coordinate to its truncated signed distance to the object surface, and its zero level set defines the reconstructed shape. We optimize θ by minimizing a joint objective that enforces geometric consistency with the tactile-derived geometry while encouraging semantic consistency through a diffusion-based prior on the class-level text description.

Tactile-based Geometric Supervision Consider reconstructing an object that is touched K times at different locations, producing tactile readings $\mathbf{T}_1, ..., \mathbf{T}_K$. As described in Sec. 3.1, we convert each tactile reading into a virtual camera observation consisting of depth maps, normal maps, and contact regions. We use these observations to impose sparse geometric constraints on the object surface represented by the SDF function f_θ. However, to impose these constraints, we need to first represent the virtual camera observations in a ray-based form compatible with our reconstruction algorithm (Fig. 2d).

Let $\mathcal{R}$ denote the union of all rays cast from the K virtual cameras $\{\mathbf{C}_K\}$, where each ray originates from a virtual camera center and passes through a pixel within the contacted region. For each ray $r \in \mathcal{R}$, the tactile-derived depth and normal maps in the virtual camera frame provide the depth observation $d(r)$ and surface normal observation $\mathbf{n}(r)$ along that ray.

To enforce geometric constraints from the tactile-derived ray observations $d(r)$ and $\mathbf{n}(r)$, we adopt the volumetric rendering formulation of Neuralangelo [18]. For each ray $r \in \mathcal{R}$, we use the SDF function f_θ along the ray to compute the predicted depth $d_\theta(r)$ and surface normal $\mathbf{n}_\theta(r)$. These computed quantities are differentiable with respect to the SDF parameters θ, allowing us to enforce geometric consistency by minimizing the discrepancy between the SDF-computed depth and normal values and the tactile-derived depth and normal values:

$$\mathcal{L}_{\text{depth}} = \mathbb{E}_{r\sim\mathcal{R}}\left[|d(r) - d_\theta(r)|\right], \qquad \mathcal{L}_{\text{normal}} = \mathbb{E}_{r\sim\mathcal{R}}\left[\|\mathbf{n}(r) - \mathbf{n}_\theta(r)\|_1\right] \quad (1)$$

Beyond supervision on these computed quantities, we incorporate two additional supervision signals following [2]. The first supervises the signed distance values near the observed surface, and the second enforces freespace constraints along the ray up to the surface. To supervise the SDF, for each ray r with a tactile-derived depth observation $d(r)$, we sample depths s within a truncated band $\mathcal{S}_r^{\text{sdf}} = [d(r) - \delta,\ d(r) + \delta]$ around the surface, where δ denotes the truncation distance. Let $\mathbf{x}(r, s)$ denote the 3D point at depth s along ray r, the SDF loss is:

$$\mathcal{L}_{\text{sdf}} = \mathbb{E}_{r\sim\mathcal{R}}\ \mathbb{E}_{s\sim\mathcal{S}_r^{\text{sdf}}}\left|f_\theta(\mathbf{x}(r, s)) - (s - d(r))\right|. \quad (2)$$

To enforce freespace constraints, we additionally sample points in $\mathcal{S}_r^{\text{fs}} = [0,\ d(r) - \delta]$, which extends from the virtual camera center to a distance δ before the surface. Since this region should be physically occupied by the tactile sensor and should remain free of objects, we penalize the predicted signed distance to be smaller than the distance δ:

$$\mathcal{L}_{\text{fs}} = \mathbb{E}_{r\sim\mathcal{R}}\, \mathbb{E}_{s\sim\mathcal{S}_r^{\text{fs}}} \left[\text{ReLU}(\delta - f_\theta(\mathbf{x}(r,s)))^2\right]. \tag{3}$$

We optimize a weighted combination of these four losses to enforce geometrical consistency with the tactile observations. In practice, the expected values of the losses are estimated by sampling a batch of rays $r \in \mathcal{R}$ during each training iteration.

Diffusion-based Prior To guide the reconstructed geometry to align with the class-level text description (e.g. "a camera"), we incorporate a diffusion prior. In particular, we use Stable Diffusion [35], a general-purpose generative model, and apply score distillation sampling (SDS) [26] to impose geometric and semantic guidance during optimization (Fig. 2c).

At each optimization step, we uniformly sample a virtual camera pose from a sphere centered at the object and render a surface normal image $\mathbf{N}_\theta$ via volumetric rendering of the SDF f_θ. Following Fantasia3D [6], we supervise the geometry using the normal map rather than an RGB image to focus on providing fine surface details. We denote $\mathbf{z}(\mathbf{N}_\theta)$ as the latent feature obtained by passing the rendered normal map $\mathbf{N}_\theta$ through the Stable Diffusion VAE encoder. The SDS gradient with respect to the SDF parameters θ is:

$$\nabla_\theta \mathcal{L}_{\text{SDS}} = \mathbb{E}_{t,\epsilon}\left[w(t)\big(\hat{\epsilon}_\phi(\mathbf{z}_t(\mathbf{N}_\theta); y, t) - \epsilon\big)\frac{\partial \mathbf{z}(\mathbf{N}_\theta)}{\partial \theta}\right] \tag{4}$$

where $\hat{\epsilon}_\phi$ is the noise predicted by Stable Diffusion, y is the text description, and ϵ is the actual noise added to the latent $\mathbf{z}$ to produce the noisy version $\mathbf{z}_t$. Additionally, t denotes the diffusion timestep, and $w(t)$ is a timestep-dependent weighting function. By backpropagating this SDS gradient, we refine the SDF parameters θ, so the reconstructed geometry matches the text description y.

Training Schedule The final optimization objective combines the tactile-defined geometric losses (Eq. (1), Eq. (2), and Eq. (3)), the diffusion-guided SDS loss (Eq. (4)), and the Eikonal regularization term. We adopt a two-stage training strategy. In the warm-up stage (steps 0–1000), optimization is driven only by tactile supervision to establish a geometry consistent with the tactile readings. In the joint refinement stage (steps 1000-7000), we optimize the model with both tactile supervision and SDS guidance, where the diffusion model operates on 64×64 rendered normal images with a batch size of 8.

3.3 Stage 2: Learning Fine Geometry of the Object

To refine geometric details beyond the coarse reconstruction, we further learn fine object geometry using DMTet, an explicit tetrahedral-grid SDF representation. In Sec. 3.2, the diffusion-based prior is constrained to apply to low-resolution rendered images of 64×64, because the underlying geometry is represented by an MLP that predicts SDF values and requires expensive per-ray field queries for

volumetric rendering. Rendering at higher resolutions is therefore computationally prohibitive, which limits the diffusion prior's ability to recover fine geometric details. In contrast, DMTet adopts a fully explicit representation. It discretizes the signed distance field onto a tetrahedral grid with vertices $\mathcal{V} = \{v_i\}_{i=1}^{N_v}$ and parameterizes the geometry by per-vertex signed distance values $\{s_i\}_{i=1}^{N_v}$ and per-vertex offsets $\{\Delta v_i\}_{i=1}^{N_v}$, where N_v is the number of grid vertices. This representation eliminates the need for MLP evaluation and per-ray field queries. The object surface can be extracted using differentiable marching tetrahedra and rendered with differentiable rasterization, which is significantly more efficient than rendering with the Neuralangelo-style SDF representation used in Sec. 3.2. Consequently, DMTet enables rendering at a much higher resolution of 512×512 with batch size 4.

Let $\phi = \{\{s_i\}, \{\Delta v_i\}\}$ denote the learnable parameters of DMTet. We initialize s_i by querying the SDF of the optimized coarse geometry from Sec. 3.2 at each grid vertex. Similar to learning the coarse geometry, we optimize ϕ by minimizing the weighted sum of the depth loss $\mathcal{L}_{\text{depth}}$, normal loss $\mathcal{L}_{\text{normal}}$, and the SDS loss $\mathcal{L}_{\text{SDS}}$. Different from stage 1, these losses are now computed through efficient differentiable rasterization instead of ray casting and MLP evaluation. In particular, the SDS loss is evaluated at a much higher rendering resolution, enabling the diffusion-based prior to recover finer geometric details. Beyond these losses, we additionally include the normal consistency loss [34] that penalizes angular deviations between adjacent vertex normals, encouraging locally smooth surface geometry:

$$\mathcal{L}_{\text{nc}} = \frac{1}{|\mathcal{E}|} \sum_{(i,j)\in\mathcal{E}} \left(1 - \cos(\mathbf{n}_i, \mathbf{n}_j)\right), \tag{5}$$

where $\mathcal{E}$ denotes the set of mesh edges, and $\mathbf{n}_i$ and $\mathbf{n}_j$ are the unit vertex normals at the endpoints of edge (i, j).

4 Experiments and Results

We evaluate the reconstruction performance of TouchAnything in simulation and compare it with baseline methods. We additionally demonstrate its qualitative performance in real-world scenarios and conduct ablation studies to examine the impact of different modality inputs. All objects were trained on a single NVIDIA A100 GPU which takes $\sim$ 1 hour for stage 1 and 40 minutes for stage 2.

4.1 Simulation Experiments

Data Collection We evaluate TouchAnything on 280 objects from ShapeNetCore.V2 [5]. We use the same object subset as TouchSDF [8], provided by its authors, to enable direct comparison with it and subsequent work [46]. The objects span six categories: bowls, bottles, cameras, jars, guitars, and mugs. The class-level text prompt associated with each category is defined respectively as

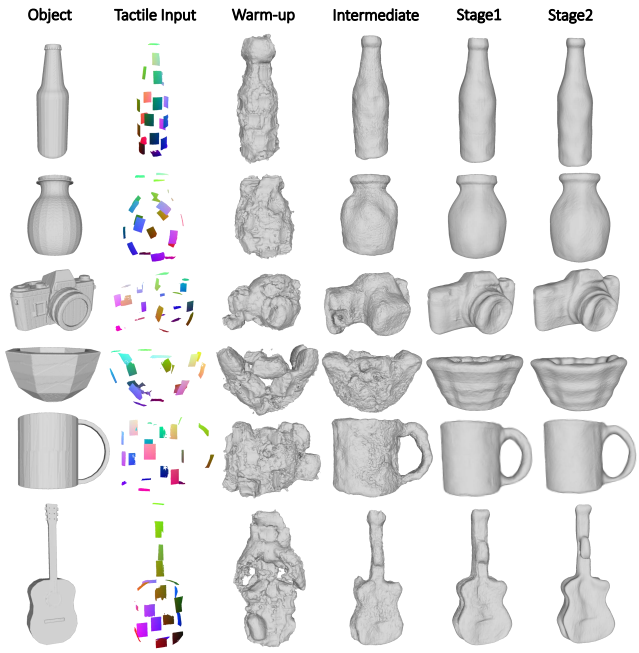

Fig. 3: Visualization of the simulation results. From left to right, we show the ground truth object, the tactile measurements, the result from the warm-up stage where only tactile observations are used, followed by intermediate, stage 1, and stage 2 reconstructions. The warm-up stage provides a good initialization of the shape before incorporating the diffusion model. We also note that the stage 1 results are close to stage 2 because the simulated objects contain limited geometric detail.

"a bowl", "a bottle", "a camera", "a jar", "a guitar", and "a mug". For each object, we generate 20 simulated tactile images as input to TouchAnything. This is achieved through a tactile simulation pipeline designed to closely approximate how a robot interacts with objects in real-world scenarios. We use Taxim [40], an example-based photorealistic tactile simulator, to produce GelSight images. For each object mesh, we uniformly sample contact locations (see Supplementary Material for details), align the contact orientation with the local surface normal, and press the object to a predefined depth. Open3D ray casting [54] is then used to compute the resulting contact depth map, which Taxim converts into a photorealistic GelSight image with a resolution of 320×240 corresponding to a sensing area of 2.0cm $\times$ 1.5cm.

Baseline Methods We compare our approach to two tactile-based 3D reconstruction methods. **TouchSDF** [8] reconstructs 3D geometry using TacTip tactile sensors and is a common baseline in tactile-based 3D reconstruction research. It represents object geometry using DeepSDF, a neural implicit representation pretrained on a specific dataset, which provides a dataset-specific

shape prior during reconstruction. **Touch2Shape** [46] uses an active tactile exploration strategy and additionally trains a diffusion model on tactile data to provide a generative shape prior during reconstruction. In contrast, our approach uses a general-purpose model guided only by minimal text descriptions, enabling open-world reconstruction of arbitrary objects.

Table 1: Quantitative Results on the simulation data. We show the Earth Mover's Distance (EMD) metric for various methods on test objects with 20 robot touches.

Category	TouchSDF [8]	Touch2Shape [46]	Ours
Bottle	0.047 ± 0.024	0.041	$\mathbf{0.035} \pm 0.020$
Bowl	0.048 ± 0.017	0.049	$\mathbf{0.039} \pm 0.010$
Camera	0.092 ± 0.043	0.056	$\mathbf{0.047} \pm 0.025$
Guitar	0.155 ± 0.087	0.064	$\mathbf{0.060} \pm 0.025$
Jar	0.071 ± 0.038	0.055	$\mathbf{0.053} \pm 0.023$
Mug	0.066 ± 0.018	0.049	$\mathbf{0.044} \pm 0.008$

Experimental Results We evaluate the reconstruction results using Earth Mover's Distance (EMD), which is widely used in touch-based reconstruction and better reflects visual quality [8]. The result is reported in Tab. 1. TouchAnything achieves better reconstruction performance across all categories compared to both baselines, highlighting our method's ability to leverage the rich geometric priors encoded in off-the-shelf 2D diffusion models. In Fig. 3, we show sample results from the warmup, stage 1 and stage 2. We note that the warmup stage provides a good initialization of the geometry before further refinement with the diffusion prior while the later incorporation of the diffusion model further improves the reconstruction quality.

4.2 Real-world Experiments with a Robot

Hardware The real-world experiments are conducted with a 6-DOF UR5e robot arm equipped with a Gelsight Mini tactile sensor. The sensor operates at a resolution of 320×240, corresponding to a sensing area of 2.0cm $\times$ 1.5cm. For real-world evaluation, we selected 14 objects in total, including 6 objects selected from the YCB dataset, 3 3D-printed objects chosen from ShapeNetCore.V2 [5] and 5 household objects. The 3D-printed objects were designed to be mounted on a cuboid base. During the experiments, all objects were rigidly fixed to the table using a dedicated mount.

Data Collection We collected the real-world data by operating the robot arm to make contact with the object. The contact poses were selected to best cover the surface of the object uniformly. A GelSight image is collected after every touch, and the pose of the sensor is calculated using the forward kinematics of the robot arm.

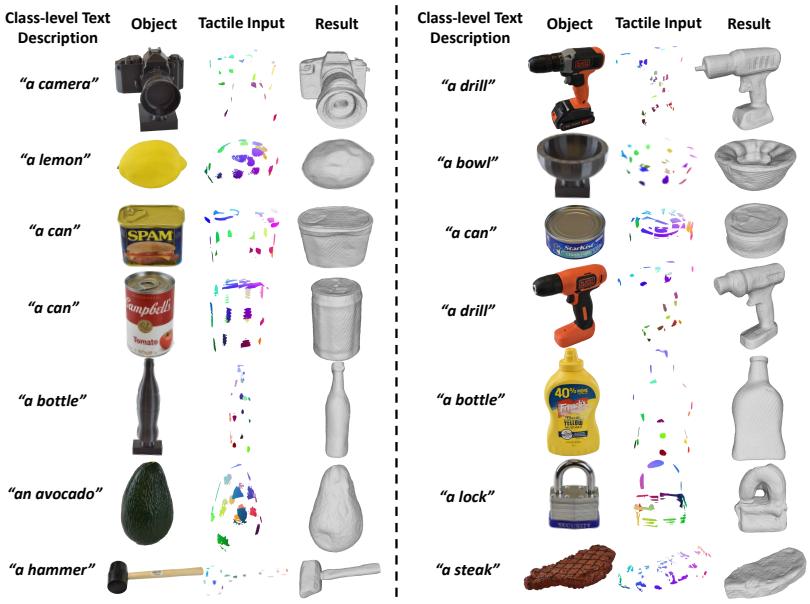

Fig. 4: Real-world reconstruction results with TouchAnything. For each object, we show the class-level text description, an image of the real object, and the 20 tactile measurements used for reconstruction. The rightmost column shows the results obtained from stage 2.

Experimental Results Fig. 4 presents the tactile input and reconstruction results for all real-world objects. For each object, 20 touches are applied uniformly across the surface. Despite the sparse tactile observations, TouchAnything achieves good reconstruction of the object geometry. For complex objects such as the camera and drill, large portions of the object remain untouched. Nevertheless, our diffusion prior correctly completes these regions with detailed shapes consistent with the semantic description. More importantly, as shown in Fig. 4, our method successfully reconstructs all tested real-world objects. This is possible because we use a general-purpose generative model rather than class-specific models. In contrast, prior methods [8, 46, 52] can typically reconstruct only objects from categories seen during training which highlights the stronger generalization capability of our approach. Fig. 5 shows the fine geometric details recovered during the stage 2 refinement step compared to the coarse geometry obtained in stage 1. Additional qualitative results are available in the supplementary material.

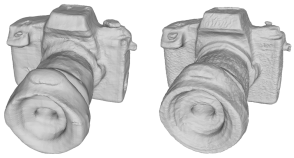

Fig. 5: Stage 1 results on the left and stage 2 result on the right. Note the finer details recovered via the refinement step.

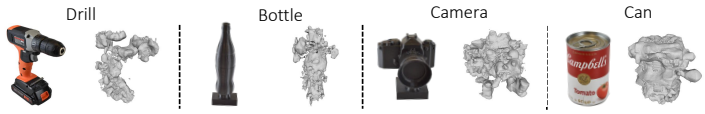

Fig. 6: Reconstructions using only tactile observations after removing the diffusion prior. The reconstructions degrade significantly.

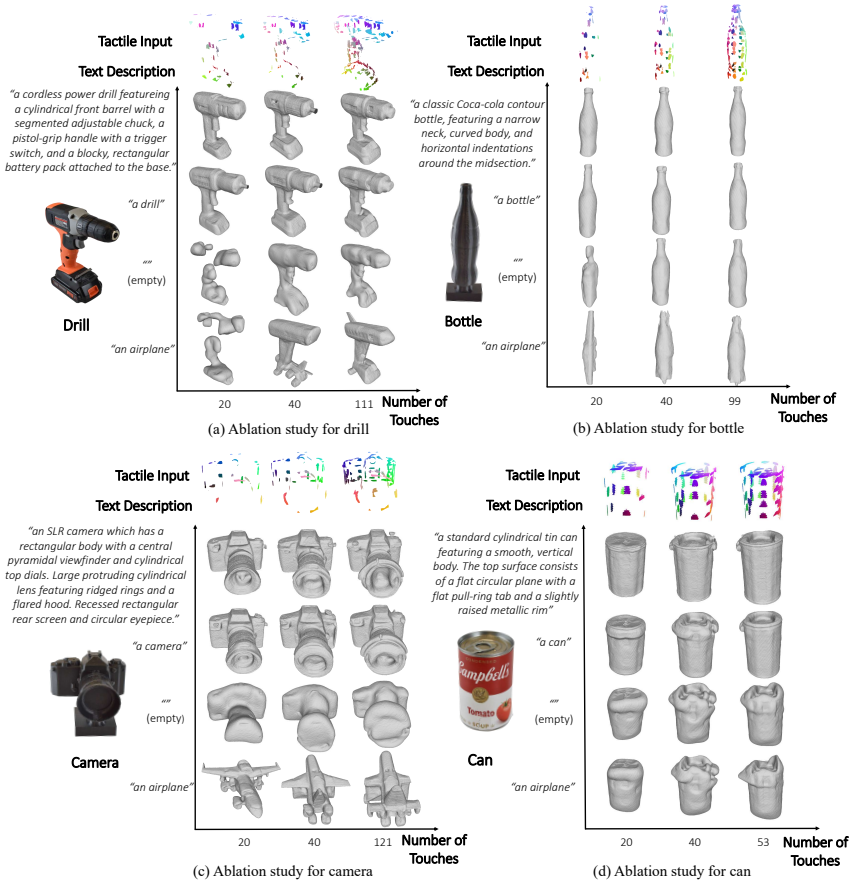

Fig. 7: We study the effect of the number of tactile measurements and the text description used to guide the diffusion prior on reconstruction quality. Columns correspond to increasing numbers of touches, while rows correspond to different text descriptions (detailed, class-level, empty, and incorrect). Our method reconstructs plausible geometries with as few as 20 touches while more informative descriptions guide the reconstruction toward more realistic shapes.

4.3 Ablation Studies

Ablation: Tactile-Only Reconstruction. Fig. 6 shows 3D reconstructions of real objects obtained using tactile observations only, after removing the diffusion prior. The reconstruction quality degrades significantly compared to our full method, highlighting the importance of the diffusion model. This behavior is expected, as tactile sensing provides only local geometric measurements. With a limited number of touches, the reconstruction problem remains highly underconstrained without the global guidance provided by the diffusion prior.

Ablation: Text Description and Touch Count. We study how the number of touches and the quality of text description affect the reconstruction performance of TouchAnything. To evaluate the effect of additional tactile supervision, we run TouchAnything using 20, 40, and all available touches. We also investigate the affect of semantic guidance by varying the text description provided to the model. Specifically, we consider four levels of semantic text guidance: an incorrect description (e.g., "an airplane"), no description (an empty string), a class-level description (e.g., "a camera"), and a highly detailed instance-specific description. Fig. 7 shows the reconstruction results of TouchAnything under different input combinations on four real-world objects: a camera, drill, cola bottle, and tomato soup can. With only 20 tactile observations, we show that a class-level text description (e.g., "a camera") is sufficient for TouchAnything to correctly complete the untouched regions, and a more detailed description does not necessarily improve reconstruction as precise geometric details are hard to specify through text. The influence of the diffusion prior is further illustrated when an incorrect description, such as "an airplane," is used, which causes the model to hallucinate structures in the unseen regions that match the incorrect semantics. Even when using an empty string as the text description, we show that the diffusion model can still reasonably infer unseen geometry based on common-sense geometric reasoning, such as connecting aligned surface patches or completing partial cylindrical structures. For completeness, we include the quantitative results (EMD) of TouchAnything under different input combinations for four real-world objects with ground-truth meshes in the supplementary material.

5 Discussion and Conclusion

We present TouchAnything, a method for reconstructing 3D object geometry from sparse tactile measurements by leveraging an off-the-shelf pretrained 2D diffusion model as a semantic and geometric prior. Our approach combines local geometric constraints derived from tactile sensing with global guidance from a diffusion prior for 3D reconstruction. As a result, it addresses the fundamentally underconstrained nature of reconstructing shapes from sparse physical contacts. Unlike prior tactile reconstruction methods that rely on category-specific training or diffusion models trained directly on tactile datasets, TouchAnything transfers geometric knowledge encoded in large-scale visual diffusion models to the

tactile domain. Our results demonstrate that this combination enables accurate and robust 3D reconstruction from only a small number of touches while generalizing to previously unseen objects in an open-world setting. Several promising directions remain for future work. One direction is to investigate the possibility of removing the reliance on text prompts entirely, enabling fully prompt-free reconstruction from tactile measurements. Additionally, our current system assumes that tactile measurements are collected passively. Future work could explore active touch strategies that guide the robot toward the most informative contact locations, enabling more efficient data collection and reconstruction.

Acknowledgements

The authors thank Ruohan Zhang and Jingyi Xiang for their help with robot setup and hardware design, thank Ruihan Gao for her discussions on tactile-inspired 3D generation, and thank Yuchen Mo for the support with computing resources. Mohamad Qadri was supported in part by ONR grant N00014-24-1-2272. This work used the Delta system at the National Center for Supercomputing Applications [award OAC 2005572] through allocation CIS240782 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program [3], which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

1. Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., Hirsh, B., Huang, S., Kalambarkar, K., Kirsch, L., Lazos, M., Lezcano, M., Liang, Y., Liang, J., Lu, Y., Luk, C., Maher, B., Pan, Y., Puhrsch, C., Reso, M., Saroufim, M., Siraichi, M.Y., Suk, H., Suo, M., Tillet, P., Wang, E., Wang, X., Wen, W., Zhang, S., Zhao, X., Zhou, K., Zou, R., Mathews, A., Chanan, G., Wu, P., Chintala, S.: PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In: 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24). ACM (Apr 2024). https://doi.org/10.1145/3620665.3640366, https://docs.pytorch.org/assets/pytorch2-2.pdf
2. Azinović, D., Martin-Brualla, R., Goldman, D.B., Nießner, M., Thies, J.: Neural rgb-d surface reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6290–6301 (June 2022)
3. Boerner, T.J., Deems, S., Furlani, T.R., Knuth, S.L., Towns, J.: ACCESS: Advancing Innovation: NSF's Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support. In: Practice and Experience in Advanced Research Computing (PEARC '23). p. 4. ACM, Portland, OR, USA (July 2023). https://doi.org/10.1145/3569951.3597559, https://doi.org/10.1145/3569951.3597559
4. Calli, B., Singh, A., Bruce, J., Walsman, A., Konolige, K., Srinivasa, S., Abbeel, P., Dollar, A.M.: Yale-cmu-berkeley dataset for robotic manipulation research. The

International Journal of Robotics Research **36**(3), 261–268 (2017). https://doi.org/10.1177/0278364917700714, https://doi.org/10.1177/0278364917700714
5. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: ShapeNet: An Information-Rich 3D Model Repository. Tech. Rep. arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago (2015)
6. Chen, R., Chen, Y., Jiao, N., Jia, K.: Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22246–22256 (2023)
7. Chibane, J., Pons-Moll, G.: Implicit feature networks for texture completion from partial 3d data. In: European Conference on Computer Vision. pp. 717–725. Springer (2020)
8. 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)
9. 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)
10. Fang, I., Shi, K., He, X., Tan, S., Wang, Y., Zhao, H., Huang, H.J., Yuan, W., Feng, C., Zhang, J.: Fusionsense: Bridging common sense, vision, and touch for robust sparse-view reconstruction. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 15798–15805. IEEE (2025)
11. Gao, R., Deng, K., Yang, G., Yuan, W., Zhu, J.Y.: Tactile dreamfusion: Exploiting tactile sensing for 3d generation. Advances in Neural Information Processing Systems **37**, 29839–29863 (2024)
12. Huang, H.J., Kaess, M., Yuan, W.: Normalflow: Fast, robust, and accurate contact-based object 6dof pose tracking with vision-based tactile sensors. IEEE Robotics and Automation Letters (2024)
13. 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)
14. Kasten, Y., Rahamim, O., Chechik, G.: Point cloud completion with pretrained text-to-image diffusion models. Advances in Neural Information Processing Systems **36**, 12171–12191 (2023)
15. 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)
16. Kung, P.C., Harisha, S., Vasudevan, R., Eid, A., Skinner, K.A.: Radarsplat: Radar gaussian splatting for high-fidelity data synthesis and 3d reconstruction of autonomous driving scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27596–27606 (2025)
17. Laine, S., Hellsten, J., Karras, T., Seol, Y., Lehtinen, J., Aila, T.: Modular primitives for high-performance differentiable rendering. ACM Transactions on Graphics **39**(6) (2020)
18. Li, Z., Müller, T., Evans, A., Taylor, R.H., Unberath, M., Liu, M.Y., Lin, C.H.: Neuralangelo: High-fidelity neural surface reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
19. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 300–309 (2023)

20. Lin, T., Qadri, M., Zhang, K., Pediredla, A., Metzler, C.A., Kaess, M.: Acoustic neural 3d reconstruction under pose drift. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 12704–12711. IEEE (2025)
21. 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)
22. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph. **41**(4) (Jul 2022). https://doi.org/10.1145/3528223.3530127, https://doi.org/10.1145/3528223.3530127
23. Ni, J., Liu, Y., Lu, R., Zhou, Z., Zhu, S.C., Chen, Y., Huang, S.: Decompositional neural scene reconstruction with generative diffusion prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6022–6033 (June 2025)
24. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5480–5490 (2022)
25. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 165–174 (2019)
26. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
27. Qadri, M., Kaess, M., Gkioulekas, I.: Neural implicit surface reconstruction using imaging sonar. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 1040–1047. IEEE (2023)
28. Qadri, M., Manchester, Z., Kaess, M.: Learning covariances for estimation with constrained bilevel optimization. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 15951–15957. IEEE (2024)
29. Qadri, M., Sodhi, P., Mangelson, J.G., Dellaert, F., Kaess, M.: Incopt: Incremental constrained optimization using the bayes tree. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 6381–6388. IEEE (2022)
30. Qadri, M., Zhang, K., Hinduja, A., Kaess, M., Pediredla, A., Metzler, C.A.: Aoneus: A neural rendering framework for acoustic-optical sensor fusion. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–12 (2024)
31. Qiu, L., Chen, G., Gu, X., Zuo, Q., Xu, M., Wu, Y., Yuan, W., Dong, Z., Bo, L., Han, X.: Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9914–9925 (2024)
32. Qu, Z., Vengurlekar, O., Qadri, M., Zhang, K., Kaess, M., Metzler, C., Jayasuriya, S., Pediredla, A.: Z-splat: Z-axis gaussian splatting for camera-sonar fusion. IEEE transactions on pattern analysis and machine intelligence **47**(9), 7255–7267 (2024)
33. Rafidashti, M., Lan, J., Fatemi, M., Fu, J., Hammarstrand, L., Svensson, L.: Neuradar: Neural radiance fields for automotive radar point clouds. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2488–2498 (2025)
34. Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., Gkioxari, G.: Accelerating 3d deep learning with pytorch3d. arXiv:2007.08501 (2020)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
36. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer International Publishing, Cham (2015)
37. Schaefer, S., Galvis, J.D., Zuo, X., Leutengger, S.: Sc-diff: 3d shape completion with latent diffusion models. arXiv preprint arXiv:2403.12470 (2024)
38. Shen, T., Gao, J., Yin, K., Liu, M.Y., Fidler, S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. Advances in Neural Information Processing Systems **34**, 6087–6101 (2021)
39. Shen, T., Gao, J., Yin, K., Liu, M.Y., Fidler, S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In: Advances in Neural Information Processing Systems (NeurIPS) (2021)
40. Si, Z., Yuan, W.: Taxim: An example-based simulation model for gelsight tactile sensors. IEEE Robotics and Automation Letters **7**(2), 2361–2368 (2022). https://doi.org/10.1109/LRA.2022.3142412
41. 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)
42. 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)
43. 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)
44. Wang, G., Chen, Z., Loy, C.C., Liu, Z.: Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9065–9076 (2023)
45. Wang, S., She, Y., Romero, B., Adelson, E.: Gelsight wedge: Measuring high-resolution 3d contact geometry with a compact robot finger. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 6468–6475 (2021). https://doi.org/10.1109/ICRA48506.2021.9560783
46. Wang, Y., Zhang, Z., Qiu, J., Sun, D., Meng, Z., Wei, X., Yang, X.: Touch2shape: Touch-conditioned 3d diffusion for shape exploration and reconstruction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5656–5665 (2025)
47. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. Advances in neural information processing systems **36**, 8406–8441 (2023)
48. Wu, R., Mildenhall, B., Henzler, P., Park, K., Gao, R., Watson, D., Srinivasan, P.P., Verbin, D., Barron, J.T., Poole, B., et al.: Reconfusion: 3d reconstruction with diffusion priors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21551–21561 (2024)
49. Yuan, W., Khot, T., Held, D., Mertz, C., Hebert, M.: Pcn: Point completion network. In: 2018 international conference on 3D vision (3DV). pp. 728–737. IEEE (2018)
50. Yuan, W., Dong, S., Adelson, E.H.: Gelsight: High-resolution robot tactile sensors for estimating geometry and force. Sensors **17**(12), 2762 (2017)

51. Yuksel, C.: Sample elimination for generating poisson disk sample sets. Comput. Graph. Forum **34**(2), 25–32 (May 2015). https://doi.org/10.1111/cgf.12538, https://doi.org/10.1111/cgf.12538
52. Zhang, H., Zhang, X., Huang, J., Feng, Z., Xiao, X.: End-to-end diffusion-based 3d object reconstruction from robotic tactile sensing. IEEE Robotics and Automation Letters **11**(2), 1434–1441 (2025)
53. Zhao, C., Sun, S., Wang, R., Guo, Y., Wan, J.J., Huang, Z., Huang, X., Chen, Y.V., Ren, L.: Tclc-gs: Tightly coupled lidar-camera gaussian splatting for autonomous driving: Supplementary materials. In: European Conference on Computer Vision. pp. 91–106. Springer (2024)
54. Zhou, Q.Y., Park, J., Koltun, V.: Open3D: A modern library for 3D data processing. arXiv:1801.09847 (2018)
55. Zou, Z., Cheng, W., Cao, Y.P., Huang, S.S., Shan, Y., Zhang, S.H.: Sparse3d: Distilling multiview-consistent diffusion for object reconstruction from sparse views. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 7900–7908 (2024)

Supplementary Material

In this supplementary material, we provide additional details regarding our methodology, experiments and results.

A Data Collection

A.1 Simulation Data Collection

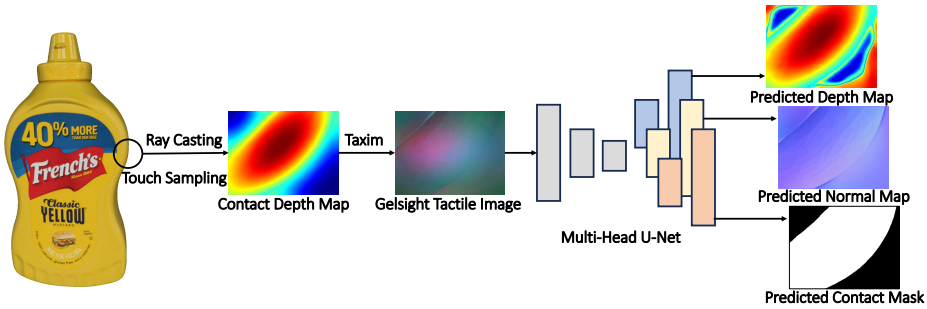

Fig. 8: A Multi-Head U-Net is implemented to derive local geometry from Gelsight Tactile Images.

Tactile image generation We propose a physically grounded simulation pipeline for tactile data generation that captures realistic contact mechanics by leveraging Taxim [40], an example-based Gelsight simulation model. The data acquisition process begins with the selection of potential contact points via Poisson Disk Sampling [51] over the object mesh; here, the surface is intentionally oversampled to ensure high-quality tactile image generation by providing a buffer against unsuitable geometries, such as overly flat regions or extreme depressions.

Once these candidates are established, the sensor poses are initialized by translating a set distance from each pre-selected contact point along the radial vector originating from the coordinate center. From these initial positions, the system perceives the local surface normals within the sensor's field of view. These sensors are subsequently aligned to the identified normals and driven into the surface at a predefined pressing depth to simulate physical contact. To synthesize the final output, a contact depth map is generated using Open3D ray casting [54], which is then transformed into a 320×240 pixel Gelsight tactile image through the Taxim framework. Finally, the samples with insufficient contact area are discarded. The sensor poses for the validated tactile contacts are stored.

For our simulation experiments conducted with ShapeNetCore.V2 [5] objects, we rescale them to 20% of their original size following TouchSDF [8]. This makes their sizes comparable with real-world sensor and objects.

Tactile image to local geometry To reconstruct the local surface geometry from raw sensory data, we implemented a Multi-Head U-Net architecture [36]. This model serves as a perception module that maps a single RGB GelSight tactile image to three distinct geometric representations: a depth map, a normal map, and a contact mask, as illustrated in Figure 8.

The network features a shared encoder to extract common tactile features, followed by three independent decoder heads tailored for each modality. We trained the model on a large-scale synthetic dataset comprising 20k tactile samples. These samples were generated using the Taxim simulation framework based on 78 YCB objects with varying mesh resolutions (16k and 64k). Each tactile sample set includes a ground truth depth map, a ground truth normal map, and a ground truth contact mask.

The model is optimized end-to-end using a weighted multi-task loss function $\mathcal{L}_{total}$, which balances the convergence across different tasks:

$$\mathcal{L}_{total} = \lambda_m \mathcal{L}_{mask} + \lambda_d \mathcal{L}_{depth} + \lambda_n \mathcal{L}_{normal} \tag{6}$$

where we employ Binary Cross-Entropy (BCE) for the contact mask loss $\mathcal{L}_{mask}$, L_1 loss for the depth loss $\mathcal{L}_{depth}$, and Cosine Similarity for the normal loss $\mathcal{L}_{normal}$.

For TouchAnything, we specifically leverage the predicted depth map and contact mask to synthesize virtual camera observations. The predicted normal maps, while currently redundant for our immediate geometry-derivation stage, are generated by our multi-head architecture to provide a more comprehensive geometric prediction and to facilitate alternative tactile sensing tasks in future research.

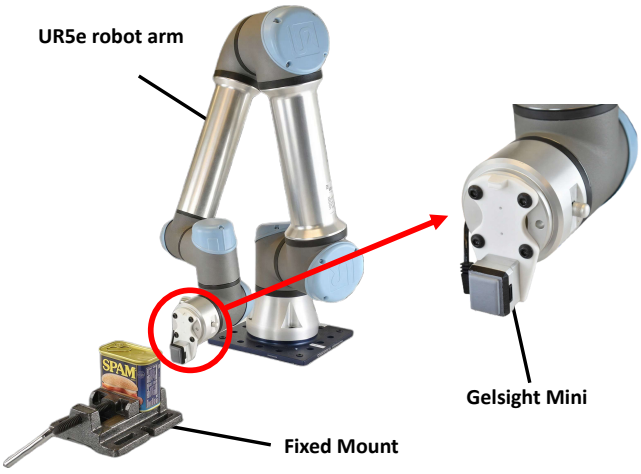

Fig. 9: Our real-world experiment hardware setup, including a UR5e robot arm, a Gelsight Mini tactile sensor and a fixed mount.

A.2 Real-world Data Collection

Our real-world data collection system includes a 6-DOF UR5e robot arm, a Gelsight Mini tactile sensor and a fixed mount on the table, as shown in Figure 9. To collect tactile data, we operate the robot arm manually and press the tactile sensor on the surface of the object. A GelSight image is recorded after every contact, and the sensor pose is calculated through the forward kinematics of the robotic arm.

B Extended Analysis of Fine Geometry Learning

In this section, we provide additional visual evidence and implementation details for the geometry refinement stage. The pipeline for the explicit DMTet [39]-based refinement is illustrated in Figure 10. Utilizing DMTet representation and differentiable mesh renderer [17], we achieve extremely fast rendering at a very high resolution. This architectural shift is the key for the diffusion model to capture high-frequency geometric details and provide rich texture to the surface.

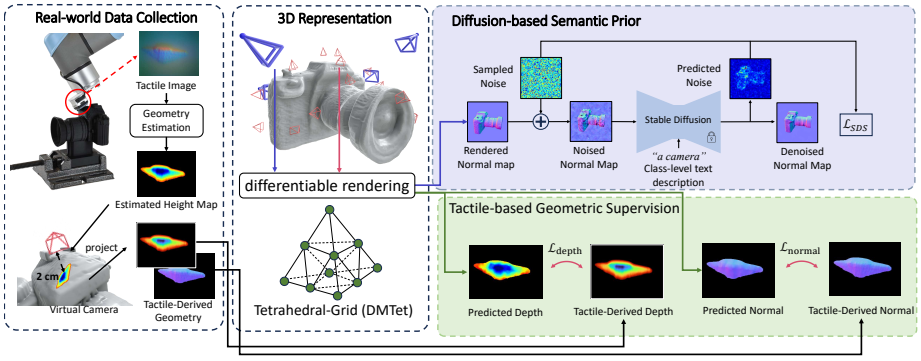

Fig. 10: Overview of the stage 2 fine geometry learning pipeline.

To demonstrate the effectiveness of our refinement strategy, we present a qualitative comparison in Figure 11. We choose some representative objects with rich surface details themselves to show how their surface texture evolves after the stage 2 geometry refinement step. While the coarse stage successfully recovers the global topology and basic structure, it often suffers from plain surface texture and lacks fine-grained details because the rendered normal images passed to the diffusion prior are of low resolution. In contrast, after our refinement, the reconstructed results all exhibit fine surface details, such as the rugged texture on the surface of the avocado, the concentric ridges on the top of the can, and the parallel fluting on the lens of the camera. These results provide compelling evidence that our refinement stage effectively recovers high-frequency surface details, leading to more fine-grained object reconstructions.

Fig. 11: Qualitative comparison between the results before and after fine geometry learning. We choose some representative objects to show how their surface texture evolves after the stage 2 geometry refinement.

C Quantitative Evaluation for Real-world Reconstruction Results

We present the quantitative metrics for the real-world reconstruction results of selected objects with ground truth meshes. These include the potted meat can, the tomato soup can, the tuna can and the mustard bottle from the YCB dataset [4] which have ground truth meshes provided by high-resolution scanner, and the printed camera model, the printed bottle model and the printed bowl model, whose ground truth meshes are from the ShapeNet-Core.V2 [5] dataset.

They are evaluated using Earth Mover's Distance (EMD). The result is reported in Figure 12.

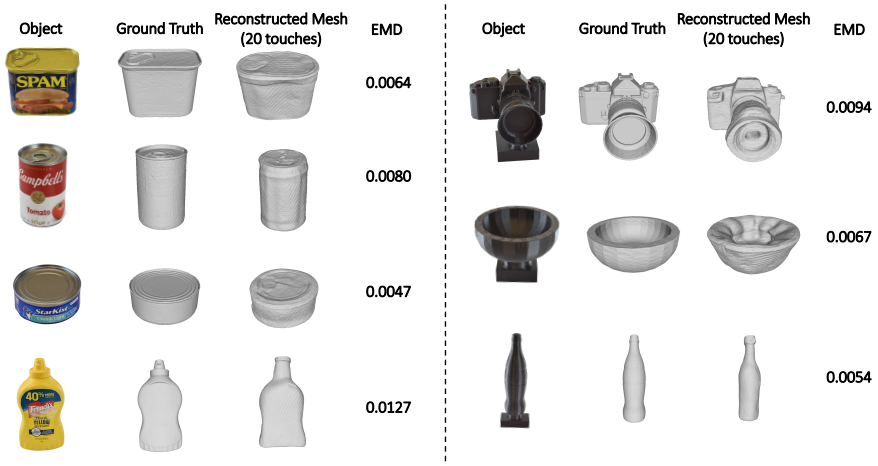

Fig. 12: Quantitative evaluation for real-world results.

For our ablation study, to further analyze the impact of tactile sample density and semantic guidance, we provide comprehensive Earth Mover's Distance (EMD) metrics for four representative real-world objects: a 3D-printed bottle, a 3D-printed camera, a tomato soup can, and a potted meat can. As shown in Table 2, we evaluate the reconstruction error across varying numbers of robot touches and four levels of text prompts (p_1: incorrect, p_2: empty, p_3: class-level, and p_4: detailed). The details of text prompts used for each real-world object is listed in Table 3.

The results highlight two key insights regarding the synergy between tactile evidence and semantic priors. First, the quality of the text prompt plays an important role in reconstruction quality. Across most test cases, the class-level (p_3) and detailed (p_4) descriptions consistently outperform the incorrect (p_1) or empty (p_2) prompts. This demonstrates that a correct semantic prior is essential for the diffusion model to resolve the inherent ambiguity of sparse tactile points and to complete untouched regions with category-specific geometric information.

Second, we observe that reconstructions using 20 or 40 touches occasionally yield lower EMD values than those using the full tactile dataset. We attribute this to the non-uniform quality of tactile measurements in dense sampling. During full-scale data collection, the sensor inevitably interacts with highly complex or sharp features, such as the thin metal ring edge on a can or the narrow edges of a camera lens hood. Due to the elastic deformation of the tactile sensor, depth estimation for these thin structures often suffers from a low-pass filtering effect, resulting in thickened artifacts that deviate from the ground truth mesh. In contrast, with fewer but cleaner samples used as local constraints, our pipeline

Table 2: Detailed EMD results for the ablation study, including four representative real-world objects: a 3D-printed bottle, a 3D-printed camera, a tomato soup can, and a potted meat can.

(a) 3D-Printed Bottle

Prompt \ Touches	20	40	99
p_1	0.0105	0.0067	0.0075
p_2	0.0058	0.0058	0.0080
p_3	0.0057	0.0052	0.0066
p_4	0.0089	0.0050	0.0056

(b) 3D-Printed Camera

Prompt \ Touches	20	40	121
p_1	0.0180	0.0190	0.0221
p_2	0.0104	0.0071	0.0112
p_3	0.0098	0.0095	0.0109
p_4	0.0097	0.0094	0.0100

(c) Tomato Soup Can

Prompt \ Touches	20	40	53
p_1	0.0099	0.0090	0.0090
p_2	0.0089	0.0088	0.0100
p_3	0.0077	0.0082	0.0087
p_4	0.0079	0.0069	0.0079

(d) Potted Meat Can

Prompt \ Touches	20	40	91
p_1	0.0092	0.0080	0.0082
p_2	0.0086	0.0086	0.0081
p_3	0.0062	0.0069	0.0074
p_4	0.0064	0.0071	0.0080

leverages the powerful geometric priors guided by accurate prompts (p_3, p_4) to reconstruct plausible structures.

Table 3: Details of text prompts (p_1 to p_4) used for each real-world object in the ablation study.

Object	ID	Text Prompt Content
3D-Printed Bottle	p_1	an airplane
	p_2	[Empty String]
	p_3	a bottle
	p_4	a classic Coca-cola contour bottle, featuring a narrow neck, curved body, and horizontal indentations around the midsection
3D-Printed Camera	p_1	an airplane
	p_2	[Empty String]
	p_3	a camera
	p_4	an SLR camera which has a rectangular body with a central pyramidal viewfinder and cylindrical top dials. Large protruding cylindrical lens featuring ridged rings and a flared hood. Recessed rectangular rear screen and circular eyepiece.
Tomato Soup Can	p_1	an airplane
	p_2	[Empty String]
	p_3	a can
	p_4	a standard cylindrical tin can featuring a smooth, vertical body. The top surface consists of a flat circular plane with a flat pull-ring tab and a slightly raised metallic rim
Potted Meat Can	p_1	an airplane
	p_2	[Empty String]
	p_3	a can
	p_4	a rectangular cuboid can with heavily rounded vertical edges. The top surface features a raised lip and a flat central plane. An integrated loop-shaped pull-tab attached rests flat on the top surface.

D Implementation Details

D.1 3D Representation Implementation Details

We implement TouchAnything using PyTorch [1] with 16-bit mixed precision. All models are optimized using the Adam optimizer. We set the learning rate to 1×10^{-2} for the hash grid and 1×10^{-3} for the SDF network and DMTet parameters. We employ the pretrained Stable Diffusion v2.1-base model [35] as the SDS backbone. The optimization is split into two distinct stages as follows.

Stage 1: Coarse Geometry Optimization For the implicit geometry in Stage 1, we represent the SDF using a multi-resolution hash-grid encoder followed by a shallow MLP, following the implementation of Instant-NGP [22] and Neuralangelo [18]. The encoder uses a Progressive Hash Grid [18] with 16 levels, 2 features per level, and a maximum hash table size of 2^{20}. The base resolution is set to 16 with a per-level growth scale of approximately 1.45. The levels are incrementally unlocked starting from level 8 and updates every 400 steps from step 1000. The SDF network is implemented as Vanilla MLPs with one hidden layer of 64 neurons and ReLU activations.

We adopt the NeuS volume renderer [30] for stage 1. For tactile-derived supervision, we sample with 16384 rays in a batch, and sample 512 points along each ray to compute the geometric constraints. For diffusion-based guidance, we sample 8 random camera views in a batch and input the rendered normal maps to the diffusion model.

In Stage 1, we make the normal loss weight λ_{normal} gradually increase to its target value (from 0.025 to 1.0 in simulation, and from 0.1 to 4.0 in real-world experiments) following a linear schedule over the first 6,000 steps. This progressive strategy prevents the normal supervision from disrupting the global shape optimization during the initial phase, allowing the model to prioritize coarse geometry establishment via depth constraints before shifting focus toward the refinement of fine-grained surface details in the later stages.

Stage 2: Fine Geometry Refinement Stage 2 utilizes an explicit tetrahedral grid (DMTet) with a resolution of 256, initialized from the optimized Stage 1 SDF. For surface extraction we use marching tetrahedra where the extraction threshold is set to -0.03 for simulation objects and 0.0 for real-world objects.

We use nvdiffrast [17] as the differentiable renderer for stage 2. For tactile-derived supervision, we sample min(number of touches, 32) sets of tactile-derived images to compute the geometric constraints. For diffusion-based guidance, we sample 4 random camera views in a batch and input the rendered normal maps to the diffusion model.

D.2 Multi-Head U-Net Implementation Details

The geometry estimation module for simulated tactile data is implemented as a Multi-Head U-Net [36] designed to map a GelSight tactile image $T \in \mathbb{R}^{3\times H\times W}$ to

three distinct geometric representations. The network follows an encoder-decoder paradigm with a shared feature extractor and task-specific branches.

Shared Encoder The encoder consists of a series of blocks designed to extract hierarchical tactile features. It begins with an initial double convolution (*DoubleConv*) layer that maps the input to 64 channels. This is followed by four downsampling stages (*Down*), each utilizing a 2×2 max-pooling operation followed by a *DoubleConv* block. The channel dimensions increase progressively as $\{64, 128, 256, 512, 1024\}$, effectively capturing both local texture and global contact structure. Each *DoubleConv* block consists of two 3×3 convolutions, each followed by Batch Normalization and ReLU activation.

Multi-Task Decoders To reconstruct local geometry, the shared latent features are passed into three independent decoder branches:

- **Depth Decoder:** Predicts a single-channel depth map $\hat{D} \in \mathbb{R}^{1\times H\times W}$.
- **Normal Decoder:** Predicts a three-channel surface normal map $\hat{N} \in \mathbb{R}^{3\times H\times W}$. A L_2 normalization layer is applied to the output to ensure the predicted vectors remain on the unit sphere.
- **Mask Decoder:** Predicts a single-channel contact mask $\hat{M} \in \mathbb{R}^{1\times H\times W}$ representing the probability of physical contact at each pixel.

Each decoder branch mirror the encoder's depth using four upsampling blocks (*Up*). We employ bilinear interpolation followed by a *DoubleConv* block to reduce the channel dimension while concatenating skip connections from the corresponding encoder stages. Finally, a 1×1 convolution (*OutConv*) is used to project the feature maps to the required output channels.

Implementation Details The model is trained using the AdamW optimizer with a learning rate of 1×10^{-3} and a weight decay of 1×10^{-2}. We employ a cosine annealing learning rate scheduler. In our implementation, the weights for the losses mentioned in Euqation (6) are set to $\lambda_d = 0.1$, $\lambda_n = 1.0$, and $\lambda_m = 5.0$.

Tactile Perception Network Architecture The detailed architectural parameters of the Multi-Head U-Net are summarized in Table 4. The encoder is shared across all tasks, while three independent decoder branches are utilized for depth, surface normal, and contact mask prediction.

Table 4: Detailed architecture of the Multi-Task Tactile Perception U-Net. C_{in} and C_{out} denote the number of input and output channels, respectively.

Part	Layer	Type	C_{in}	C_{out}	Configuration
Encoder	`inc`	DoubleConv	3	64	Conv 3×3, Pad 1
	`down1`	Down	64	128	MaxPool 2×2 + DoubleConv
	`down2`	Down	128	256	MaxPool 2×2 + DoubleConv
	`down3`	Down	256	512	MaxPool 2×2 + DoubleConv
	`down4`	Down	512	512	MaxPool 2×2 + DoubleConv
Decoders*	`up1`	Up	1024	256	Bilinear Up + Cat + DoubleConv
	`up2`	Up	512	128	Bilinear Up + Cat + DoubleConv
	`up3`	Up	256	64	Bilinear Up + Cat + DoubleConv
	`up4`	Up	128	64	Bilinear Up + Cat + DoubleConv
	`out`	OutConv	64	N_{task}	Conv 1×1

*Decoders for Depth ($N_{task} = 1$), Normal ($N_{task} = 3$), and Mask ($N_{task} = 1$) share the same architecture but have independent weights.