FusionSense: Bridging Common Sense, Vision, and Touch for Robust Sparse-View Reconstruction

Irving Fang[1,*], Kairui Shi[1,*], Xujin He[1,*], Siqi Tan[1], Yifan Wang[1], Hanwen Zhao[1], Hung-Jui Huang[2], Wenzhen Yuan[3], Chen Feng[1,✉], Jing Zhang[1,✉]

https://ai4ce.github.io/FusionSense/

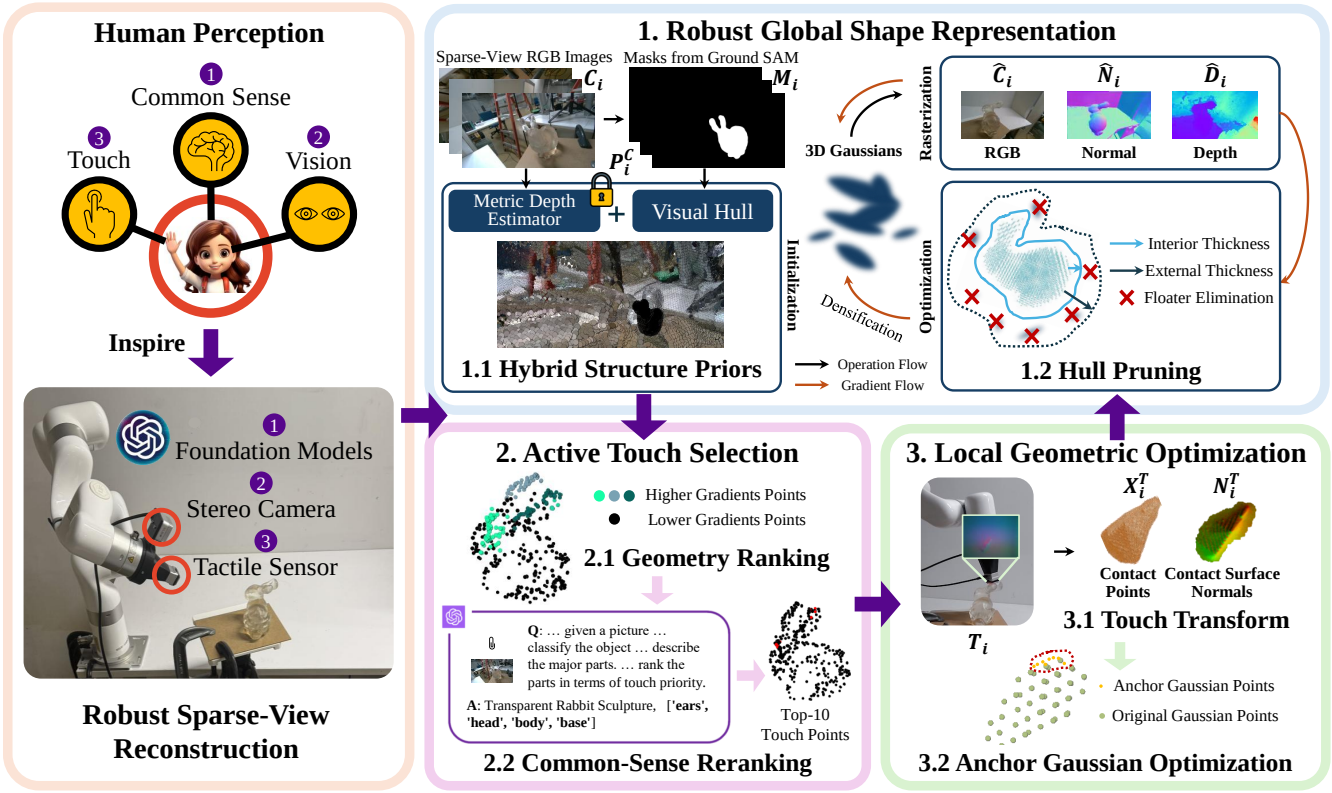

Fig. 1. **Overview of FusionSense.** Inspired by human perception, **FusionSense** integrates common sense from foundation models with *sparse-view* data from both vision and touch through 3D Gaussian Splatting, enabling efficient and robust 3D reconstruction of a robot's surroundings. Our proposed system features three core modules: (i) robust global shape representation, (ii) active touch point selection on the object, and (iii) local geometric optimization.

***Abstract*— Humans effortlessly integrate common-sense knowledge with sensory input from vision and touch to understand their surroundings. Emulating this capability, we introduce *FusionSense*, a novel 3D reconstruction framework that enables robots to fuse priors from foundation models with highly sparse observations from vision and tactile sensors. FusionSense addresses three key challenges: (i) How can robots efficiently acquire robust global shape information about the surrounding scene and objects? (ii) How can robots strategically select touch points on the object using geometric and common-sense priors? (iii) How can partial observations such as tactile signals improve the overall representation of the object? Our framework employs 3D Gaussian Splatting as a core representation and incorporates a hierarchical optimization strategy involving global structure construction, object visual hull pruning and local geometric constraints. This advancement results in fast and robust perception in environments with traditionally challenging objects that are transparent, reflective, or dark, enabling more downstream manipulation or navigation tasks. Experiments on real-world data suggest that our framework outperforms previously state-of-the-art sparse-view methods. All code and data are open-sourced on the project website.**

[1]New York University, Brooklyn, NY 11201, USA
[2]Carnegie Mellon University, Pittsburgh, PA 15289, USA
[3]University of Illinois, Urbana-Champaign, Champaign, IL 60606, USA
*Equal contributions.
✉ Corresponding authors {cfeng, jz6676}@nyu.edu. The work was supported in part through NSF grants 2024882, 2152565, and 2238968.

I. Introduction

Humans exhibit an extraordinary ability to perceive their surroundings by seamlessly integrating common-sense knowledge, vision, and touch, even when presented with sparse or incomplete views [1]. Common-sense reasoning helps bridge gaps in sensory data, vision offers a broad understanding of the environment, and touch provides fine-grained information about texture and material properties through direct physical interaction [2]. This synergy between cognitive and sensory inputs inspires more intuitive and efficient robotic perception in complex environments [3, 4].

Despite recent advances, current robotic perception systems have yet to fully harness the multimodal capabilities

that humans naturally employ. Emerging techniques like 3D Gaussian Splatting (3DGS) [5] show potential for flexible and efficient 3D reconstruction of intricate structures. However, vision-based methods [6, 7], especially those dependent on *sparse-view observations* [8], continue to face challenges such as occlusion, suboptimal lighting conditions, and difficult surfaces like transparent [9], reflective [10], or dark objects [11]. Approaches such as [12] leverage pre-trained models like DeepSDF [13] for shape completion, yet they still struggle with objects possessing unique geometries or intricate details. Conversely, high-resolution optical tactile sensors [14, 15] can overcome these limitations through direct physical interaction with high-resolution sensing, yet they have a limited sensing range. For example, the reinforcement learning strategy in [16] takes a cobot 1,631 touches to fully explore the surface of a banana in the YCB dataset [17], which has a surface area of only 216 cm^2. Furthermore, while multimodal methods combining visual and tactile data have shown promise for improving object perception and 3D reconstruction, passive touch strategies often significantly increase the number of actions needed [11, 18].

To overcome these limitations, we present **FusionSense**, a novel 3D reconstruction framework that integrates priors from foundation models with sparse observations from both vision and tactile sensors. At the core of our framework is 3D Gaussian Splatting, which provides an efficient and scalable means to represent the environment. In this framework, surface normal supervision is highlighted to enrich both global and local geometric details [19, 20, 21, 22]. Specifically, FusionSense is built upon three key modules: (**i**) **Robust Global Shape Representation**, where hybrid structure priors are introduced to initialize geometry and ensure multi-view consistency alongside a hull-pruning constraint to guide optimization for both the scene and the object; (**ii**) **Active Touch Selection**, prioritizing points with high gradients during optimization that indicate complex structures, while incorporating common-sense knowledge from foundation models for decision-making; and (**iii**) **Local Geometric Optimization**, where new anchor Gaussians are added to guide fine detail optimization, with geometric normals supervised by the high-resolution tactile feedback provided by the GelSight sensor.

These innovations lead to the following key contributions:

1) We propose a novel 3D reconstruction framework for scenes and objects that fuses priors from foundation models with *sparse observations from visual and tactile sensors*, exploiting the unique strengths of each modality. We also develop an active touch strategy driven by geometric and common-sense cues, enhancing perceptual granularity with fewer robot actions. This framework can handle objects that are traditionally challenging for 3D reconstruction, such as transparent, reflective, or dark objects.
2) We propose a novel hierarchical optimization strategy designed for 3DGS. This strategy incorporates object hull pruning to guide the optimization process and introduces anchor Gaussians at the local level, supervised by surface normals captured from tactile signals, to refine fine-grained details. Our work is the first to natively incorporate tactile signals into 3DGS.
3) We deploy our algorithm on a real robot, demonstrating its competitive ability to reconstruct surroundings with challenging objects under highly sparse observations.

II. Related Work

A. Visuo-Tactile Robot Perception

Roboticists have long been exploring tactile sensing. In 1984, Bajcsy and Goldberg [23] had already explored tactile surface reconstruction with primitive tactile sensors. Recently leap in tactile technology [14, 24] enabled great progress in object classification [25, 26], deformable object manipulation [27, 28, 29], industrial insertion [30, 31], etc.

In tactile-only reconstruction, researchers often employ an active strategy to select touch points due to the limited coverage area of tactile sensors. Many of them [32, 33] chose the Gaussian Process Implicit Surface as the representation for the shape, of which the derived uncertainty drives the selection strategy. Matsubara *et al.* [34] used end-effector travel distance as another constraint to accelerate the procedure, while Shahidzadeh *et al.* [16] sidestepped Gaussian Process and utilized reinforcement learning for an exploration policy.

In visuo-tactile works, visual signals can provide a rough global shape of the object, greatly reducing the number of touches and enabling passive touch strategies. Swann *et al.* [11] and Suresh *et al.* [18] employed a grid-like, exhaustive touch strategy. Smith *et al.* [35] only considered touch patch at the grasping spot. For active touch, Smith *et al.* [36] learned a strategy in simulation, while Björkman *et al.* [37] and Wang *et al.* [38] again employed uncertainty in Gaussian Process. A key observation is that the uncertainty in the Gaussian Process usually comes from a lack of visual signal on certain parts. The part itself may be otherwise unremarkable. However, our active strategy also focuses on the geometrically complicated and fine-grained parts. In addition, our method is the first to employ the state-of-the-art Gaussian Splatting [5] method instead of a simple baseline method for initial reconstruction, the first to employ multiple foundation models, and the first to efficiently fuse tactile signal natively into Gaussian Splatting [5], unlike in TouchGS [11] where the tactile signal is still incorporated via Gaussian Process Implicit Surface [39].

B. Gaussian Splatting for 3D Reconstruction

Gaussian Splatting (3DGS) [5] is a fast and efficient method for 3D reconstruction and radiance field rendering, representing scenes with Gaussian primitives to preserve continuous volumetric properties while enabling rapid optimization and real-time rendering. DN-Splatter [20] improves upon this by introducing geometric normal supervision, enhancing geometric accuracy, particularly in textureless regions, but its performance is limited under sparse-view observations. GaussianObject [8] is designed for object reconstruction in sparse-view settings and uses the visual hull to initialize the object point cloud, though its optimization process differs from ours. TouchGS [11] optimizes 3DGS using GPIS

results [39] derived from dense tactile observations (e.g., 632 touches), making it inefficient for robotic applications. Inspired by these methods, we present the first framework to fuse common sense and sparse observations from both vision and touch using Gaussian primitives without being limited by the number of touches.

III. Method

A. Problem Formulation

Our goal is to represent a previously unseen scene, S, using a set of differentiable 3D Gaussian primitives, $\mathcal{G} = \{\boldsymbol{G}_k : \boldsymbol{p}_k, \boldsymbol{q}_k, \boldsymbol{s}_k, o_k, \boldsymbol{c}_k\}_{k=1}^{K}$. The geometry of each Gaussian $\boldsymbol{G}_k$ is parameterized by its center $\boldsymbol{p}_k \in \mathbb{R}^3$, rotation quaternion $\boldsymbol{q}_k \in \mathbb{R}^3$, and scaling vector $\boldsymbol{s}_k \in \mathbb{R}^3$. The appearance parameters include opacity $o_k \in \mathbb{R}$, and color $\boldsymbol{c}_k \in \mathbb{R}^3$. Rendering a new view is achieved by projecting 3D Gaussians into 2D space from the camera's perspective. The resulting 2D Gaussians are depth-sorted globally and then alpha-composited using the discrete volume rendering equation to compute the final pixel colors, $\hat{C}$, depth estimation, $\hat{D}$, and Normal estimation, $\hat{N}$ [20]:

$$\hat{C} = \sum_{k \in N} c_k \alpha_k T_k, \hat{D} = \sum_{k \in N} d_k \alpha_k T_k, \hat{N} = \sum_{k \in N} n_k \alpha_k T_k, \tag{1}$$

where $T_k = \prod_{j=1}^{k-1}(1-\alpha_j)$ is the accumulated transmittance at pixel location p and α_k is the blending coefficient for a Gaussian with center μ_k in screen space:

$$\alpha_k = o_k \cdot \exp\left(-\frac{1}{2}(p-\mu_k)^{\mathrm{T}} \Sigma_k^{-1}(p-\mu_k)\right). \tag{2}$$

In particular, we are interested in a challenging object O that may be transparent, reflective, or dark. We aim to reconstruct it in some digital representation (in this case, 3D Gaussian primitives) as close to the original as possible.

To this end, we collect the following sparse observations from vision and tactile sensors and will fuse them with priors from foundation models:

- **RGB $\boldsymbol{C}_i$** and **depth $\boldsymbol{D}_i$ images**, and their **pose $\boldsymbol{P}_i^{\mathrm{C}}$** in the world frame with the following dimensions: $\boldsymbol{C}_i \in \mathbb{R}^{1280\times720\times3}$, $\boldsymbol{D}_i \in \mathbb{R}^{1280\times720\times1}$, $\boldsymbol{P}_i^{\mathrm{C}} \in SE(3)$.
- **Tactile** signal $\boldsymbol{T}_i$ and its **pose** $\boldsymbol{P}_i^{\mathrm{T}}$ in the world frame with the following dimension: $\boldsymbol{T}_i \in \mathbb{R}^{240\times320\times3}$, $\boldsymbol{P}_i^{\mathrm{T}} \in SE(3)$.

Note that tactile signals are saved as RGB images to be processed later. Also, note that the RGB and depth images are aligned so they share the same pose.

B. Method Overview

Our method can be divided into three modules:

1) Robust Global Shape Representation: This module leverages hybrid geometry priors and object hull pruning to optimize a global 3D representation, denoted as $\mathcal{G}$, that contains the scene and the object of interest $\boldsymbol{O}$. The hybrid geometry prior combines monocular depth estimates $\overline{\boldsymbol{D}}_i$ [40], camera poses $\boldsymbol{P}_i^{\mathrm{C}}$, and visual hull results $\overline{\boldsymbol{O}}$ [41] to produce an initial representation $\mathcal{G}'$. During optimization, hull pruning eliminates floating artifacts and ensures a clean representation of the initial reconstructed object $\boldsymbol{O}'$, derived from both $\overline{\boldsymbol{O}}$ and the global shape $\mathcal{G}$. $\mathcal{G}'$ is supervised with RGB $\boldsymbol{C}_i$, depth $\boldsymbol{D}_i$, and normal priors $\overline{\boldsymbol{N}}_i$ from [19].

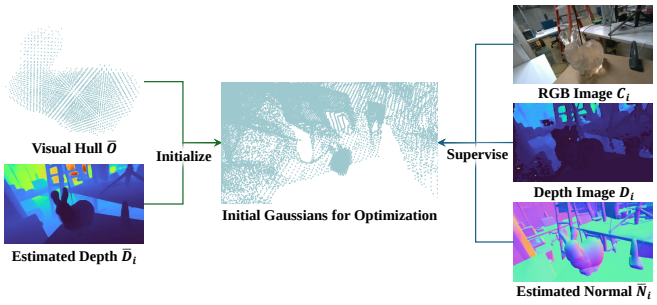

Fig. 2. We use visual hull and estimated depth to initial our Gaussians and use RGB, depth, and estimated normal to supervise the training.

2) Active Touch Selection: This module proposes touch points $\boldsymbol{t}_i$ on $\boldsymbol{O}'$ where tactile feedback is needed. The robot then collects tactile signals $\boldsymbol{T}_i$ at these points. It consists of two sub-modules:

- A geometry-focused module that ranks points in regions with high gradients in 3DGS, indicating intricate structures or discrepancies between splatting and the image.
- A common-sense-driven module that utilizes large vision and language models (VLMs) to rerank points from the previous module, integrating common-sense knowledge from VLMs to enhance decision-making.

3) Local Geometric Optimization: This module takes in $\boldsymbol{T}_i$ and the contact masks $\boldsymbol{M}_i^{\mathrm{T}}$, surface normals $\boldsymbol{N}_i^{\mathrm{T}}$, and contact points $\boldsymbol{X}_i^{\mathrm{T}}$ [18], introducing an anchor Gaussian optimization strategy. Anchor Gaussians $\boldsymbol{G}^{\mathrm{T}}$ are initialized from $\boldsymbol{X}_i^{\mathrm{T}}$ and further refined using $\boldsymbol{N}_i^{\mathrm{T}}$ and global context. By integrating tactile signals $\boldsymbol{T}_i$ into the global representation $\mathcal{G}$, this module refines local geometric details.

A summary of symbols can be found in Table I.

C. Robust Global Shape Representation

To obtain a robust global representation $\mathcal{G}$, we introduce hybrid structure priors and hull pruning strategies. Specifically, we adopt a variant of 3DGS [20] that incorporates surface normal supervision.

Hybrid structure priors are employed to ensure multi-view consistency. First, we estimate the coarse geometry $\overline{\boldsymbol{O}}$ of the target object using a visual hull [41], which is constructed by combining camera poses $\boldsymbol{P}_i^{\mathrm{C}}$ and segmented silhouettes $\boldsymbol{M}_{\mathrm{o}}$ extracted via Grounded SAM 2 [42]. This method is independent of surface appearance, resilient to challenging materials, and key to our success with objects that are otherwise challenging to traditional reconstruction methods. Next, we acquire the surrounding coarse geometry $\overline{\boldsymbol{S}}$ using monocular depth priors $\overline{\boldsymbol{D}}_i$ from depth foundation model Metric3D v2 [40], along with the corresponding camera poses $\boldsymbol{P}_i^{\mathrm{C}}$. These hybrid structure priors are fused by applying distance thresholds τ_{d} to integrate $\overline{\boldsymbol{O}}$ and $\overline{\boldsymbol{S}}$ to produce the initial global representation $\mathcal{G}'$, which contains the initial reconstructed $\boldsymbol{O}'$, for further optimization.

During the subsequent optimization process, we design hull pruning to remove the floaters in the exterior region

TABLE I

LIST OF IMPORTANT SYMBOLS AND THEIR DEFINITIONS

Symbol	Definition	Symbol	Definition	Symbol	Definition	Symbol	Definition
S	Unseen scene	$\mathcal{G}$	3DGS primitives	$\boldsymbol{C}_i, \boldsymbol{D}_i$	Camera RGB, depth images	$\boldsymbol{M}_i^T$	Contact mask
O	Object to reconstruct	$\mathcal{G}'$	Initial 3DGS primitives	$\boldsymbol{P}_i^{\mathrm{C}}$	Camera poses	$\boldsymbol{N}_i^T$	Contact surface normals
$\boldsymbol{O}$	Recon'd object in 3DGS	$\hat{C}$	3DGS pixel values	$\overline{\boldsymbol{D}}_i, \overline{\boldsymbol{N}}_i$	Estimated depth, normals	$\boldsymbol{X}_i^T$	Contact points
$\boldsymbol{O}'$	Initial recon'd object in 3DGS	$\hat{D}$	3DGS depth	$\boldsymbol{T}_i$	Tactile signals	$\boldsymbol{G}^T$	Anchor 3DGS Primitives
$\overline{\boldsymbol{O}}$	Object visual hull	$\hat{N}$	3DGS normals	$\boldsymbol{P}_i^{\mathrm{T}}$	Tactile sensor poses	$\boldsymbol{p}_k, \boldsymbol{q}_k$	Primitives pose
$\overline{\boldsymbol{O}}_{\mathrm{s}}$	Shell around visual hall	$o_k, \boldsymbol{c}_k$	Primitives opacity, color	$\boldsymbol{M}_o$	SAM2 object mask	$\boldsymbol{s}_k$	Primitives scaling factor

outside the hull $\overline{\boldsymbol{O}}$. Gaussian primitives are particularly sensitive to these floaters, as they can slow convergence and result in suboptimal outcomes, especially when dealing with sparse observations. Hull pruning is achieved by introducing a thin shell $\overline{\boldsymbol{O}}_{\mathrm{s}}$ surrounding the hull $\overline{\boldsymbol{O}}$. $\overline{\boldsymbol{O}}_{\mathrm{s}}$ is defined by two thickness parameters: an interior thickness $t_{\mathrm{s}}^{\mathrm{i}}$ and an external thickness $t_{\mathrm{s}}^{\mathrm{e}}$. In our setup, $t_{\mathrm{s}}^{\mathrm{i}}$ is set to be larger than 5 mm, corresponding to the voxel grid resolution of the visual hull, while $t_{\mathrm{s}}^{\mathrm{e}}$ is empirically set to 2 cm. Then, similar to [20], we utilize RGB $\boldsymbol{C}_i$ and depth $\boldsymbol{D}_i$ images and normal $\bar{N}_i$ estimated from normal foundation model DSINE [43] to supervise $\hat{C}, \hat{D}, \hat{N}$ from our $\mathcal{G}'$, as seen in Fig. 2.

D. Active Touch Selection

An active touch strategy with geometric and common-sense cues can reduce the number of touches needed.

1) Geometry: We capitalize on the design of the original 3DGS [5] that high gradients at each Gaussian primitive indicate rapid changes in spatial features and larger discrepancies between the rendering generated by the Gaussians and the image, which means a need for further optimization.

Given the objective $\boldsymbol{O}'$ for tactile interaction, we use the densification mean value from Sec. III-C as the gradient threshold τ_g to select some Gaussian primitives $\boldsymbol{G}'_k$. Next, we apply DBSCAN [44] algorithm to cluster $\boldsymbol{G}'_k$, filtering out outliers. Then, all the selected Gaussians are ranked based on mean gradient values in its cluster, forming a ranking $\mathcal{R}_G$.

2) Common Sense: The sub-module provides another ranking $\mathcal{R}_C$ by leveraging common-sense knowledge from vision-language models (VLMs).

First, we randomly select one RGB image from the captured images $\boldsymbol{C}_i$ and prompt GPT-4-o [45] with the image and descriptive text to obtain a classification label and a list of relevant part names, along with a ranking $\mathcal{R}_p$ of the parts based on priority in touching.

To ground this common-sense ranking $\mathcal{R}_p$ to the object $\boldsymbol{O}'$, we utilize a zero-shot open-vocabulary part segmentation model, PartSLIP [46]. Based on a textual prompt of parts names, PartSLIP classifies each point of an extracted point cloud from $\boldsymbol{O}'$ into a specific part as in Fig. 3. From Sec. III-C, we know the coordinates of every Gaussian in $\boldsymbol{G}'_k$ and every point in the extracted point cloud from $\boldsymbol{O}'$ as can be seen in Fig. 3. We also know each point's ranking in $\mathcal{R}_p$. So, we iterate through $\boldsymbol{G}'_k$, assigning every Gaussian a rank based on the closest point in the point cloud, thus forming another ranking $\mathcal{R}_C$ for selected Gaussians $\boldsymbol{G}'_k$.

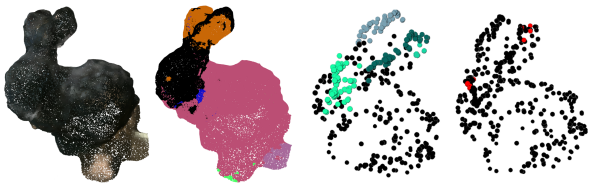

Fig. 3. **(1)** Point Cloud Extracted from $\boldsymbol{O}'$. **(2)** Part Segmentation from PartSLIP. **(3)** High Gradient Gaussians. **(4)** 10 Selected Touch Points $\boldsymbol{t}_i$.

We then sort $\boldsymbol{G}'_k$ based on $\mathcal{R}_C$ first and then $\mathcal{R}_G$, ensuring that even if PartSLIP in the second sub-module fails to work properly, we still have a reasonable, geometrically sensible touch sequence $\boldsymbol{t}_i$, as seen in Fig. 3

E. Local Geometric Optimization

This module enhances local geometric detail by transforming the tactile signal $\boldsymbol{T}_i$ into contact masks $\boldsymbol{M}_i^{\mathrm{T}}$, surface normals $\boldsymbol{N}_i^{\mathrm{T}}$, and contact points $\boldsymbol{X}_i^{\mathrm{T}}$ [18]. We then introduce anchor Gaussian optimization to integrate the tactile signals $\boldsymbol{T}_i$ into the global representation $\mathcal{G}$.

Given a tactile signal $\boldsymbol{T}_i$ as an RGB image, because the tactile sensor is made of a gel patch that has consistent optical properties across all its surface, we can calculate a mapping between surface gradients $(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y})$ and the RGB value at a given location (r, g, b, x, y) based on photometric stereo [47]. In practice, this mapping is acquired by pressing a ball with a known radius on the gel patch and recording the corresponding tactile image. Then, a multi-layer perceptron can be trained after manually labeling the deformation caused by the ball. Assuming that the gel patch is the zero level surface of a scalar field $f(x, y) - z$, the contact surface normal $\boldsymbol{N}_i^{\mathrm{T}}$ can be derived from the surface gradient as $(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, -1)$ [14]. Applying a Poisson solver to integrate the surface gradients gives us a depth map of the gel patch's shape. We can then acquire a contact mask $\boldsymbol{M}_i^{\mathrm{T}}$ and, consequentially, contact points $\boldsymbol{X}_i^{\mathrm{T}}$ with a depth threshold.

Contact points $\boldsymbol{X}_i^{\mathrm{T}}$ are added as anchor Gaussians $\boldsymbol{G}^{\mathrm{T}}$ due to the scale difference between $\boldsymbol{T}_i$ and visual signals $\boldsymbol{C}_i$. Treating $\boldsymbol{X}_i^{\mathrm{T}}$ as ground truth, we fix the center $\boldsymbol{p}^{\mathrm{T}}$ and opacity o^{T} of $\boldsymbol{G}^{\mathrm{T}}$, while optimizing the rotation $\boldsymbol{q}^{\mathrm{T}}$, scale $\boldsymbol{s}^{\mathrm{T}}$, and color $\boldsymbol{c}^{\mathrm{T}}$. Notably, we apply Gaussian normal supervision directly to $\boldsymbol{G}^{\mathrm{T}}$ instead of normal images. This allows the integration of $\boldsymbol{G}^{\mathrm{T}}$ into $\mathcal{G}$, combining local surface normals $\boldsymbol{N}_i^{\mathrm{T}}$ with global information $\boldsymbol{C}_i$, $\boldsymbol{D}_i$ to refine the final geometry.

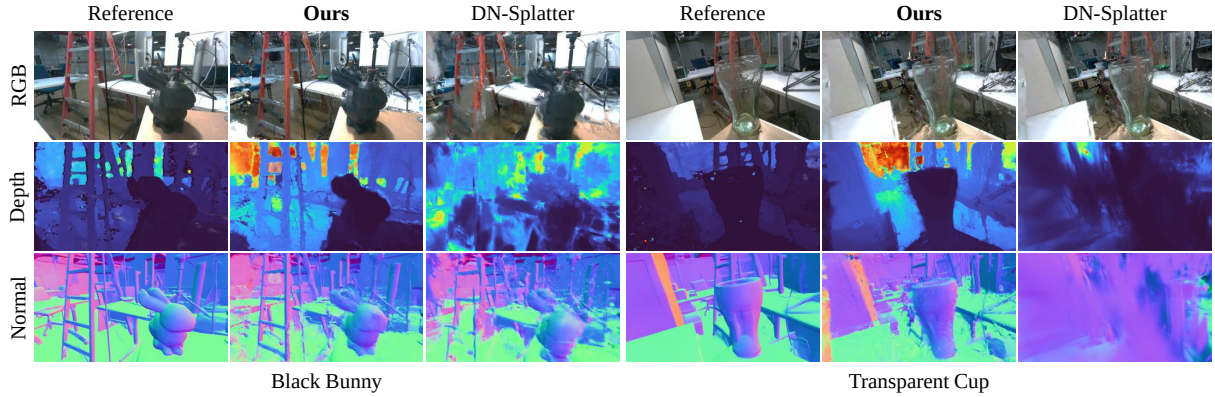

Fig. 4. **Qualitative comparisons on novel view synthesis, depth estimation, and normal estimation under sparse observations.** The comparison presents results from scenes with two challenging objects: a black bunny and a transparent Coca-Cola cup. Comparisons are made between (i) the reference (ground truth RGB images, depth images from a RealSense camera, and normal estimates generated by the DSINE monocular normal foundation model [43]), (ii) the proposed **FusionSense** framework, and (iii) the DN-Splatter approach. Using sparse observations—9 views and 10 tactile contacts—**FusionSense** achieves higher image fidelity, more precise depth, and normal estimations compared to DN-Splatter [20], which relies on 9 views.

IV. Experiment

A. Experiment Setup

1) Robot and Sensor: Our experiments are conducted using a GelSight Mini tactile sensor for acquiring tactile signal $\boldsymbol{T}_i$, an Intel RealSense D405 for acquiring RGB $\boldsymbol{C}_i$ and depth $\boldsymbol{D}_i$ images and a 6 DOF UFactory xArm 6 cobot with 0.1-millimeter repeatability. The camera and tactile sensor are mounted to the robot's end-effector with a 3D printed mount, so we know the dimensions and can easily calculate accurate transformations between each sensor and the end-effector, and therefore, the robot base, which also serves as the origin of our world frame.

2) Data Collection and Challenging Objects: We conduct real-world robot experiments and collect data from surrounding scenes featuring four challenging objects, comparable to the tactile baseline [11] in quantity and difficulty. The challenging objects are categorized into (1) 3D-printed challenging dark objects and (2) non-3D-printed objects. All the 3D-printed objects are printed with a high-precision Formlabs resin printer [48] before they are painted to achieve high transparency, reflection, or darkness.

3) Comparison Methods and Metrics: We compared our method with three representative 3D Gaussian Splatting (3DGS) approaches: (i) DN-Splatter for scene and object reconstruction, (ii) GaussianObject [8] for sparse-view object reconstruction, and (iii) TouchGS[1] for visual-tactile integration. We evaluate **scene reconstruction** using standard novel view synthesis metrics such as **PSNR** and **SSIM** [5]. To assess **object reconstruction** quality, we calculate the **Chamfer Distance (CD)** between the reconstructed surface point clouds and ground truth point clouds, which are downsampled from the CAD models of the 3D-printed objects.

4) Implementation Details: The proposed Gaussian Splatting training approach is implemented using PyTorch and `gsplat` library [49]. To produce a well-performing visual hull, up to 5% of the selected grid points may not be observed in the mask image. All models are trained for 15,000 iterations, and densification begins at 800 iterations. The touch patches are added at iteration 1,000 to the Gaussian scene as anchor points.

[1]Since their method requires a large number of touches (e.g., 632), we use the results reported in [11] for comparison. We compare their 8-view against our 5-view and 152-view against our 9-view.

TABLE II

Quantitative Comparisons of Novel View Synthesis with Varying Input Views for Scene Reconstruction

Method	5 Views		9 Views	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
DN-Splatter [20]	13.56	**0.58**	13.32	0.55
TouchGS[1] [11]	11.75	0.47	15.51	**0.66**
FusionSense (Ours)	**16.09**	0.57	**18.83**	0.65

TABLE III

Quantitative Comparisons of Novel View Synthesis with Varying Input Views for Object Reconstruction

Method	5 Views		9 Views	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
GaussianObject [8]	11.38	**0.53**	12.73	0.57
DN-Splatter [20]	14.33	0.43	13.75	0.46
FusionSense (Ours)	**18.33**	0.51	**19.84**	**0.58**

B. Experiment Results

We evaluate our method's performance in appearance and geometry representation against other representative methods. The results demonstrate that our method unifies the advantages of the other three approaches and performs better in rendering novel views and geometric representations.

Our baseline DN-splatter [20] incorporates depth and normal supervision into 3DGS training to enhance reconstruction quality under dense views. But as shown in Fig 4 and Fig 5, DN-splatter struggles with sparse views. Metrics from Table II and III further validate this observation. With insufficient views, DN-splatter's randomly initialized point cloud will likely be stuck in local minima during optimization, leading to artificial floating artifacts between object and scene and an incoherent representation of the target object. As a result, the extracted surface point cloud would not represent the target object well, reflected by its poor Chamfer Distance (CD) performance shown in Table IV.

In contrast, GaussianObject [8] initialize GS training using Visual Hull [41]. However, its RGB-only supervision lacks

TABLE IV

CHAMFER DISTANCE ↓ (MM) FOR OBJECT RECONSTRUCTION

Method	# Touches	5 Views	9 Views
DN-Splatter [20]	0	0.237	0.192
TouchGS [11]	632	**0.023**	N/A
FusionSense (Ours)	10	0.025	**0.022**

depth and normal supervision, resulting in poor depth estimations and Gaussian gradients of the surface. Moreover, it is particularly difficult to achieve good results with RGB training alone when the target has challenging material. As Table III shows, GaussianObject performs the worst in reconstructing the four challenging objects.

In contrast, our approach uses segmented foreground and background points as seed points, providing a better initialization of the approximate positions and coarse shapes of objects and scenes. We further enhance our approach by integrating RealSense depth data and normal priors from the foundation model for supervision. Additionally, we regularize Gaussian training through hull pruning, which removes floating Gaussian points between the object and the background. These techniques significantly reduce false occluding Gaussian points from novel perspectives, contributing to our improved rendering and geometric results.

Besides, considering that we are more concerned with the quality of the target object than the entire scene, we use masks generated by Grounded SAM 2 [42] to select the pixels corresponding to the target object. This allows us to calculate object-specific PSNR and SSIM metrics.

TouchGS [11] achieves better CD with rich tactile sensing information from 632 touches by fusing touch points with the implicit surface to generate extra depth and uncertainty information. However, this approach strongly depends on the number and positioning of touches, which fails when the touch information is sparse. In contrast, our method circumvents this reliance by focusing only on empirically complex regions identified by the large language model. As shown in Table IV, we achieve a competitive geometrical result with just 10 touches under a sparse view scenario, as opposed to their 632 touches.

C. Ablation Study

1) Hull Pruning: As mentioned in Sec. III-C, hull pruning is a major modification that enables our framework. As shown in Table V, our framework without hull pruning suffers worse results in both scene and object reconstruction tasks. Without highly accurate depth supervision, many outliers will be generated during 3DGS training. While the RealSense camera performs well for close-range scenes, it struggles with distant scenes and object edges. Additionally, depth estimates produced by large models like Metric3D v2 [40] may perform well in a single viewpoint, but they often have incorrect scaling and cannot be accurately projected into a 3D model within an entire 3D scene. Therefore, hull pruning is particularly important. It effectively prevents the Gaussian points of the target object from becoming blurred or losing edge clarity due to background interference, all while not disrupting the rendering of the surrounding scene.

TABLE V

ABLATION STUDY FOR HULL PRUNING AND TOUCH STRATEGY

Data Description	Setting	PSNR↑	CD↓
Black Bunny	w/o Hull Pruning	20.05	0.0305
	Random Touch	20.50	0.0183
	Active Touch	**20.87**	**0.0176**
Transparent Bunny	w/o Hull Pruning	21.46	0.0485
	Random Touch	21.51	**0.0247**
	Active Touch	**21.65**	0.0267

Fig. 5. Rendering Results Using 5 Views.

2) Touch Strategy: Active touch strategy is another major design in our pipeline. As shown in Table V, our strategy yields slightly better results, although not across the board. There are several possible reasons: (1) the number of touches is limited, and the sizes of our objects are small in the scene. Thus, it is not easy to distinguish the effectiveness of different strategies. (2) Our first module in Sec. III-C gives unexpectedly outstanding precise results, leaving relatively little room for improvement for different touch strategies.

V. CONCLUSION, LIMITATION, AND FUTURE WORK

In this work, by fusing visual, tactile, and common-sense information, we propose a novel framework that significantly improves the state-of-the-art of scene and object reconstruction regarding challenging objects. Accompanying this framework, we propose a hierarchical optimization strategy designed for 3DGS that utilizes visual hull pruning and is the first to natively incorporate tactile signals into 3DGS without limiting touch numbers.

Meanwhile, we realize the limitations of our experiments and methods. Due to time constraints, we are not able to demonstrate the reconstruction framework's impact on downstream robotics tasks. In addition, our touch selection strategy can use more design and experiments. Currently, its effectiveness remains marginal, and the investigation into it remains limited. Another limitation lies in the process of extracting point cloud and mesh from our trained Gaussian primitives. The fine-grained geometrical tactile details cannot be fully extracted due to its tiny scale that cannot be fully sampled during the level-set extraction approach. To handle an extensive range of multi-scaled geometrical details from scene to tactile, novel strategies need to be developed.

Currently, tactile patches are acquired through teleoperated robot control, but an automated, servoing-based method could significantly increase the number of touch interactions.

References

[1] F. Hutmacher, "Why is there so much more research on vision than on any other sensory modality?" *Frontiers in psychology*, vol. 10, p. 481030, 2019. 1

[2] H. B. Helbig and M. O. Ernst, "Optimal integration of shape information from vision and touch," *Experimental brain research*, vol. 179, pp. 595–606, 2007. 1

[3] P. Allen, "Surface descriptions from vision and touch," in *Proceedings. 1984 IEEE International Conference on Robotics and Automation*, vol. 1, 1984, pp. 394–397. 1

[4] N. Navarro-Guerrero, S. Toprak, J. Josifovski, and L. Jamone, "Visuo-haptic object perception for robots: an overview," *Autonomous Robots*, vol. 47, no. 4, pp. 377–403, 2023. 1

[5] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–14, 2023. 2, 4, 5

[6] Z. Yu, T. Sattler, and A. Geiger, "Gaussian opacity fields: Efficient and compact surface reconstruction in unbounded scenes," *arXiv preprint arXiv:2404.10772*, 2024. 2

[7] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "Dust3r: Geometric 3d vision made easy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 697–20 709. 2

[8] C. Yang, S. Li, J. Fang, R. Liang, L. Xie, X. Zhang, W. Shen, and Q. Tian, "Gaussianobject: Just taking four images to get a high-quality 3d object with gaussian splatting," *arXiv preprint arXiv:2402.10259*, 2024. 2, 5

[9] Y. Cai, J. Qiu, Z. Li, and B. Ren, "Neuralto: Neural reconstruction and view synthesis of translucent objects," *ACM Transactions on Graphics (TOG)*, vol. 43, no. 4, pp. 1–14, 2024. 2

[10] Y. Liu, P. Wang, C. Lin, X. Long, J. Wang, L. Liu, T. Komura, and W. Wang, "Nero: Neural geometry and brdf reconstruction of reflective objects from multiview images," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–22, 2023. 2

[11] A. Swann, M. Strong, W. K. Do, G. S. Camps, M. Schwager, and M. Kennedy III, "Touch-gs: Visual-tactile supervised 3d gaussian splatting," *arXiv preprint arXiv:2403.09875*, 2024. 2, 5, 6

[12] M. Comi, Y. Lin, A. Church, A. Tonioni, L. Aitchison, and N. F. Lepora, "Touchsdf: A deepsdf approach for 3d shape reconstruction using vision-based tactile sensing," *IEEE Robotics and Automation Letters*, 2024. 2

[13] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 165–174. 2

[14] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017. 2, 4

[15] S. Wang, Y. She, B. Romero, and E. Adelson, "Gelsight wedge: Measuring high-resolution 3d contact geometry with a compact robot finger," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 6468–6475. 2

[16] A.-H. Shahidzadeh, S. J. Yoo, P. Mantripragada, C. D. Singh, C. Fermüller, and Y. Aloimonos, "Actexplore: Active tactile exploration on unknown objects," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 3411–3418. 2

[17] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The ycb object and model set: Towards common benchmarks for manipulation research," in *2015 international conference on advanced robotics (ICAR)*, 2015, pp. 510–517. 2

[18] S. Suresh, Z. Si, J. G. Mangelson, W. Yuan, and M. Kaess, "Shapemap 3-d: Efficient shape mapping through dense touch and vision," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 7073–7080. 2, 3, 4

[19] G. Bae and A. J. Davison, "Rethinking inductive biases for surface normal estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9535–9545. 2, 3

[20] M. Turkulainen, X. Ren, I. Melekhov, O. Seiskari, E. Rahtu, and J. Kannala, "Dn-splatter: Depth and normal priors for gaussian splatting and meshing," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025. 2, 3, 4, 5, 6

[21] K. Mazur, G. Bae, and A. Davison, "SuperPrimitive: Scene reconstruction at a primitive level," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[22] X. Cao and T. Taketomi, "Supernormal: Neural surface reconstruction via multi-view normal integration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 581–20 590. 2

[23] K. Goldberg and R. Bajcsy, "Active touch and robot perception," in *Cognition and Brain Theory*, 1984. 2

[24] T. P. Tomo, A. Schmitz, W. K. Wong, H. Kristanto, S. Somlor, J. Hwang, L. Jamone, and S. Sugano, "Covering a robot fingertip with uskin: A soft electronic skin with distributed 3-axis force sensitive elements for robot hands," *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 124–131, 2018. 2

[25] R. Corcodel, S. Jain, and J. van Baar, "Interactive tactile perception for classification of novel object instances," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 9861–9868. 2

[26] A. Schmitz, Y. Bansho, K. Noda, H. Iwata, T. Ogata, and S. Sugano, "Tactile object recognition using deep learning and dropout," in *2014 IEEE-RAS International Conference on Humanoid Robots*, 2014, pp. 1044–1050. 2

[27] A. Burns, S. Xiang, D. Lee, L. Jackel, S. Song, and V. Isler, "Look and listen: A multi-sensory pouring network and dataset for granular media from human demonstrations," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 2519–2524. 2

[28] M. Kaboli, K. Yao, and G. Cheng, "Tactile-based manipulation of deformable objects with dynamic center of mass," in *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, 2016, pp. 752–757. 2

[29] Y. She, S. Wang, S. Dong, N. Sunil, A. Rodriguez, and E. Adelson, "Cable manipulation with a tactile-reactive gripper," *The International Journal of Robotics Research*, vol. 40, no. 12-14, pp. 1385–1401, 2021. 2

[30] S. Dong and A. Rodriguez, "Tactile-based insertion for dense box-packing," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 7953–7960. 2

[31] S. Dong, D. K. Jha, D. Romeres, S. Kim, D. Nikovski, and A. Rodriguez, "Tactile-rl for insertion: Generalization to objects of unknown geometry," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 6437–6443. 2

[32] Z. Yi, R. Calandra, F. Veiga, H. van Hoof, T. Hermans, Y. Zhang, and J. Peters, "Active tactile object exploration with gaussian processes," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 4925–4930. 2

[33] N. Jamali, C. Ciliberto, L. Rosasco, and L. Natale, "Active perception: Building objects' models using tactile exploration," in *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, 2016, pp. 179–185. 2

[34] T. Matsubara and K. Shibata, "Active tactile exploration with uncertainty and travel cost for fast shape estimation of unknown objects," *Robotics and Autonomous Systems*, vol. 91,

pp. 314–326, 2017. 2

[35] E. Smith, R. Calandra, A. Romero, G. Gkioxari, D. Meger, J. Malik, and M. Drozdzal, "3d shape reconstruction from vision and touch," *Advances in Neural Information Processing Systems*, vol. 33, pp. 14 193–14 206, 2020. 2

[36] E. Smith, D. Meger, L. Pineda, R. Calandra, J. Malik, A. Romero Soriano, and M. Drozdzal, "Active 3d shape reconstruction from vision and touch," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 064–16 078, 2021. 2

[37] M. Björkman, Y. Bekiroglu, V. Högman, and D. Kragic, "Enhancing visual perception of shape through tactile glances," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2013, pp. 3180–3186. 2

[38] S. Wang, J. Wu, X. Sun, W. Yuan, W. T. Freeman, J. B. Tenenbaum, and E. H. Adelson, "3d shape perception from monocular vision, touch, and shape priors," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 1606–1613. 2

[39] O. Williams and A. Fitzgibbon, "Gaussian process implicit surfaces," in *Gaussian Processes in Practice*, 2006. 2, 3

[40] M. Hu, W. Yin, C. Zhang, Z. Cai, X. Long, H. Chen, K. Wang, G. Yu, C. Shen, and S. Shen, "Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation," *arXiv preprint arXiv:2404.15506*, 2024. 3, 6

[41] A. Laurentini, "The visual hull concept for silhouette-based image understanding," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 16, no. 2, pp. 150–162, 1994. 3, 5

[42] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan *et al.*, "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024. 3, 6

[43] G. Bae and A. J. Davison, "Rethinking inductive biases for surface normal estimation," 2024. [Online]. Available: https://arxiv.org/abs/2403.00712 4, 5

[44] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96. AAAI Press, 1996, p. 226–231. 4

[45] OpenAI *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023. 4

[46] M. Liu, Y. Zhu, H. Cai, S. Han, Z. Ling, F. Porikli, and H. Su, "Partslip: Low-shot part segmentation for 3d point clouds via pretrained image-language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 21 736–21 746. 4

[47] R. Woodham, "Photometric method for determining surface orientation from multiple images," *Optical Engineering*, vol. 19, 1992. 4

[48] Formlabs, "High resolution sla and sls 3d printers for professionals," https://formlabs.com/, 2024, accessed: 2024-09-12. 5

[49] V. Ye, R. Li, J. Kerr, M. Turkulainen, B. Yi, Z. Pan, O. Seiskari, J. Ye, J. Hu, M. Tancik, and A. Kanazawa, "gsplat: An open-source library for gaussian splatting," *arXiv preprint arXiv:2409.06765*, 2024. 5