Snap-it, Tap-it, Splat-it: Tactile-Informed 3D Gaussian Splatting for Reconstructing Challenging Surfaces

Mauro Comi
University of Bristol
mauro.comi@bristol.ac.uk

Alessio Tonioni
Google

Jonathan Tremblay
NVIDIA

Max Yang
University of Bristol

Valts Blukis
NVIDIA

Yijiong Lin
University of Bristol

Nathan F. Lepora*
University of Bristol

Laurence Aitchison*
University of Bristol

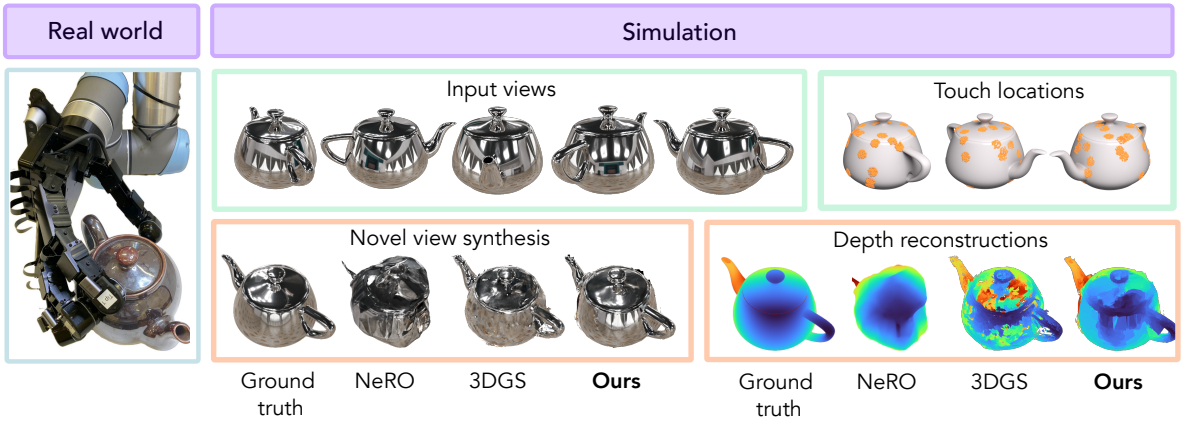

Figure 1. We combine multi-view images and tactile sensing information within a 3D Gaussian Splatting framework for accurate geometry reconstruction and novel view synthesis of challenging surfaces. On the left, a robotic arm equipped with tactile sensors interacts with an object to collect contact surface information. On the right, we show results on novel view synthesis and depth reconstruction, using a five minimal-view setting and tactile input (first row). Touches not only improve surface reconstruction, but also strengthen novel view synthesis.

Abstract

Touch and vision go hand in hand, mutually enhancing our ability to understand the world. From a research perspective, the problem of mixing touch and vision together is underexplored and presents interesting challenges. To this end, we propose Tactile-Informed 3DGS, a novel approach that incorporates contact data (local depth maps) with multi-view images to achieve surface reconstruction and novel view synthesis. Our method optimises 3D Gaussian primitives to accurately model the object's geometry at points of contact. By creating a framework that decreases the transmittance at touch locations, we achieve a refined surface reconstruction, ensuring a uniformly smooth depth map. Touch is particularly useful when considering non-Lambertian objects, such as shiny or reflective surfaces, since contemporary methods tend to fail to reconstruct specular highlights with fidelity. By combining vision and tactile sensing, we achieve more accurate geometry reconstructions with fewer images than prior methods. We conduct evaluation on objects with glossy and reflective surfaces and demonstrate improved reconstruction quality in both the virtual and real world

1. Introduction

Humans perceive the world through a multitude of senses, and often, visual perception alone is not sufficient to fully grasp the intricacy of the world. When only us-

*Authors contributed equally.

ing vision, it is possible to misunderstand shapes, materials, or scale of the world and its components. In particular, current vision-based methods often struggle with *non-Lambertian materials*, such as reflective and glossy surfaces [19], as well as scenarios where only a *limited number of views* are available. These scenarios are common in real-world settings, such as robotics manipulation, VR/AR, and 3D modeling, where physical constraints or occlusions can restrict the number of accessible viewpoints. In contrast, tactile sensing provides consistent geometric information that complements visual perception, regardless of light-dependent effects. Thus, tactile data can help to resolve visual ambiguities and can support 3D reconstruction in the context of both limited (5 views) and dense visual data. While tactile sensing provides valuable geometric information, it is limited by the fact that it only captures partial information about an object's surface, because when grasping an object, the hand rarely touches the entire surface. Given these complementary strengths and limitations, accurate 3D reconstruction in challenging settings necessitates the seamless integration of sparse observations across multiple sensing modalities (see Figure 1), which is what our work aims to achieve.

Current volume rendering techniques like 3D Gaussian Splatting (3DGS) [12] and NeRF [19] are highly effective at novel-view synthesis (NVS): the problem of generating an image from an unseen viewpoint. However, in applications such as robotic manipulation, VR/AR and 3D modelling, the problem of 3D reconstruction is often equally or even more important. While volumetric rendering techniques can be directly applied to 3D reconstruction and perform well in settings with dense viewpoints and Lambertian surfaces, they can break down in more difficult settings. Depth sensors have been proposed to aid 3D reconstruction, but they often fail in environments with highly reflective surfaces, which create discontinuities in depth measurements [28]. Recent techniques have improved vision-only 3D reconstruction for non-Lambertian surfaces by approximating the rendering equation to simulate reflections [17, 30]. However, these approaches typically require dense viewpoints and are computationally intensive (24+ hours), making them less suitable for robotic or AR/VR systems that often rely on sparse views and have strict computational time constraints. Meanwhile, these systems can provide an alternative solution by leveraging tactile information to enhance 3D reconstruction accuracy.

In that context, our key contributions are 1) to incorporate tactile sensing in a 3D Gaussian Splatting framework and 2) to show that this tactile information improves object reconstruction in settings where there is uncertainty about the object surface, such as the few-view setting, or where the surface is non-Lambertian. In that setting, we propose Tactile-Informed 3DGS, a novel multimodal interaction approach that integrates tactile sensing and vision within 3D Gaussian Splatting for challenging object reconstruction. This method operates by regularising the 3D transmittance of Gaussians around the touch locations, which directly guides the optimisation procedure with precise, direct information regarding object geometry. Acknowledging the sparse nature of tactile observations, we further refine our method by introducing an unsupervised regularisation term to smooth the predicted camera depth maps. We evaluate our method on object-centric datasets containing glossy and reflective surfaces, namely Shiny Blender [30] and Glossy Synthetic [17] datasets. We also propose a real-world dataset of a highly shiny object (metallic toaster) from which we capture a multiview set of images as well as real robot touches. Our experiments show that multimodal sensing delivers geometry reconstruction that is as good as, if not better than, advanced techniques like NeRO [17], which incorporate light scattering, all while being significantly faster. By introducing touch as an additional sensing modality, our method increases robustness and mitigates performance degradation when working with fewer viewpoints.

2. Literature Review

2.1. Tactile sensing for 3D object reconstruction

Tactile sensing for object reconstruction is an emerging research field that integrates techniques from computer vision, graphics, and robotics with high-resolution, optical-based tactile sensors [13, 14, 33]. These sensors, which capture local depth maps or marker-based images, can be leveraged to collect contact information and advance control [32] and 3D shape reconstruction through direct object interactions.

Recently we have seen the introduction of datasets that incorporate touch, such as ObjectFolder [6]. However, these datasets often fall short in providing a comprehensive solution for 3D reconstruction of challenging objects. ObjectFolder, for instance, lacks multiview images and non-Lambertian objects, limiting its applicability in real-world scenarios where reflective and glossy surfaces are common. Similarly, while Suresh *et al.* [26] introduced a dataset combining both depth cameras and tactile data, it does not offer multi-view data. To address these limitations, we create a dataset that integrates multiple views of a metallic object with corresponding tactile data.

Various approaches have been proposed to integrate tactile information in 3D reconstruction, each employing different representation techniques. Wang *et al.* [31, 33] proposed to utilise tactile exploration for object reconstruction through voxel-based representations. This method, while innovative, has limitations in capturing fine details and scaling to complex objects. These challenges are addressed by

our use of a continuous 3D representation. In a different approach, Smith *et al.* [23, 24] decoupled vision and tactile sensing processes, employing a Convolutional Neural Network (CNN) to map tactile readings into local contact surface representations. However, their experiments primarily utilised synthetic data of objects with simpler material properties (single color and Lambertian surfaces) and lacked support for novel view synthesis, which our technique inherently supports. Similarly, methods porposed to reconstruct objects shapes using touch signals [3, 22] tend to generate valuable object surfaces, but do not perform novel-view synthesis. For a comprehensive review of 3D representations coupled with tactile sensing refer to [9].

2.2. Novel-view synthesis on reflective surfaces

Novel-view synthesis from sparse observations is a key area of research within the domain of computer graphics, 3D computer vision, and robotics. Methods based on Neural Radiance Fields (NeRF) are extremely successful at synthesizing photorealistic images by modeling the volumetric density and color of light rays within a scene with a Multi-Layer Perceptron [1, 19–21]. 3D Gaussian Splatting (3DGS) [12] has been proposed as an alternative to NeRF. Instead of modelling the radiance field implicitly through a neural network, it is modelled explicitly by a large set of 3D Gaussians. This has the benefit of a simpler explicit model that offers faster training and rendering than even highly optimized NeRF implementations [20]. In this work we propose to extend the 3DGS framework to reason about non-Lambertian surfaces and touch data.

The application of volumetric rendering equations causes difficulties for both NeRFs and 3DGS in accurately representing specular and glossy surfaces. This is because it is challenging to interpolate the outgoing radiance across different viewpoints as it tends to vary significantly around specular highlights [30]. Ref-NeRF [30], which was proposed to address this limitation by replacing NeRF's view-dependent radiance parameterisation with a representation of reflected radiance, struggled with indirect lights from nearby light emitters. NeRFReN [7] models reflections by fitting an additional radiance field for the reflected world beyond the mirror, but only supports reflections on planar surfaces only. These limitations highlight the challenges faced by methods that rely uniquely on the vision sensing modality.

Within the family of methods that approximate the volume rendering equation [10, 17], perhaps the canonical example is NeRO [17], which proposes to model direct and indirect illumination separately using the split-sum approximation [11]. By directly approximating indirect specular lighting, NeRO achieves improved geometry reconstruction. The performance of NeRO in material retrieval and surface reconstruction is impressive and ranks as state-of-the-art among NeRF-based techniques, which justifies its selection as a baseline for our evaluation. However, as we show in this work, NeRO tends to generate unconvincing results when it does not have access to an extensive dataset of views (see experiments section). In contrast, GaussianShader [10] integrates a simplified shading function into 3D Gaussians, primarily for novel view synthesis. It derives depth directly from Gaussian positions, unlike our volume rendering approach. This fundamental variation in depth calculation, essential for obtaining the 3D geometry of objects, makes GaussianShader's 3D point clouds not directly comparable with our approach and other benchmarks, resulting in GaussianShader's poor performance on standard geometry reconstruction metrics.

A contemporaneous work, Touch-GS [27], was released in the same month as our study. While this work does integrate visual and touch information in the context of 3DGS, it is otherwise very different. In particular, this work uses dense views and extensive tactile data covering the entire object surface, which significantly differs from our setting that operates effectively with minimal views and sparse touch data.

3. Preliminaries

3.1. 3D Gaussian Splatting

3D Gaussian Splatting [12] is an explicit radiance field technique for modeling 3D scenes as a large set of 3D Gaussians. Specifically, a scene is parameterised as a set of N anisotropic 3D Gaussians, each defined by a mean vector $\boldsymbol{\mu} \in \mathbb{R}^3$, a covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{3\times3}$, an opacity $\alpha \in \mathbb{R}^1$, and a view-dependent color vector $\mathbf{c} \in \mathbb{R}^m$ encoded as spherical harmonics. These parameters are optimised during the training process through backpropagation. The influence of a Gaussian on a point $\mathbf{x} \in \mathbb{R}^3$ is determined by:

$$f(\mathbf{x}; \boldsymbol{\Sigma}, \boldsymbol{\mu}) = \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right). \quad (1)$$

During the optimisation process, the scene is rendered to a 2D plane by calculating each Gaussian's projected covariance $\boldsymbol{\Sigma}' \in \mathbb{R}^{2\times2}$ and position $\boldsymbol{\mu}' \in \mathbb{R}^2$, following:

$$\boldsymbol{\Sigma}' = \mathbf{JW}\boldsymbol{\Sigma}\mathbf{W}^T\mathbf{J}^T, \quad (2)$$

where $\mathbf{W}$ is the projective transformation and $\mathbf{J}$ is the Jacobian of the affine transformation. The covariance matrix $\boldsymbol{\Sigma}$, along with other parameters, is optimised to ensure a positive semi-definite matrix through decomposition:

$$\boldsymbol{\Sigma} = \mathbf{RSS}^T\mathbf{R}^T. \quad (3)$$

where $\mathbf{S}$ is a scaling matrix and $\mathbf{R}$ is a rotation matrix. The 2D mean vector $\boldsymbol{\mu}'$ results from perspective projection

$\text{Proj}(\boldsymbol{\mu}|\mathbf{E}, \mathbf{K})$, with $\mathbf{E}$ and $\mathbf{K}$ being the camera's extrinsic and intrinsic matrices. Consequently, the pixel color $I(\mathbf{p})$ in the image space is computed by summing the contributions of all Gaussians, and accounting for their color, opacity, and influence:

$$I(\mathbf{p}) = \sum_{i=1}^{N} f(\mathbf{p}; \boldsymbol{\mu}'_i, \boldsymbol{\Sigma}'_i)\ \alpha_i\ \mathbf{c}_i \prod_{j=1}^{i-1}(1 - f(\mathbf{p}; \boldsymbol{\mu}'_j, \boldsymbol{\Sigma}'_j)\ \alpha_j). \tag{4}$$

Similarly, the depth is computed as:

$$D(\mathbf{p}) = \sum_{i=1}^{N} f(\mathbf{p}; \boldsymbol{\mu}'_i, \boldsymbol{\Sigma}'_i)\ \alpha_i\ d_i \prod_{j=1}^{i-1}(1 - f(\mathbf{p}; \boldsymbol{\mu}'_j, \boldsymbol{\Sigma}'_j)\ \alpha_j), \tag{5}$$

where d_i is the depth of the i-th Gaussian obtained by projecting $\boldsymbol{\mu}_i$ onto the image frame.

The optimisation process starts by initialising the 3D Gaussians using Structure-from-Motion (SfM) point clouds. It then proceeds to optimise the learnable parameters by minimising the photometric loss between predicted and ground truth images. During the entire procedure, Gaussians are adaptively densified or pruned based on heuristics on their opacity and mean gradient changes.

4. Methodology

This section outlines our approach to reconstructing 3D objects by integrating visual and tactile data through 3D Gaussian Splatting. Our methodology follows two primary stages. The first stage involves the generation of initial point clouds from local depth maps, while the second step consists of optimising and regularising a Gaussian representation to accurately model the object's surface.

4.1. Gaussian initialisation and optimisation

The optimisation process starts with the extraction of a pointcloud from COLMAP, where each point's position and associated color serve as the initial mean and color attributes for a *first* set of Gaussians G_c. The view-dependent color is encoded using spherical-harmonics, which are essential in modeling the specular and glossy highlights. This initial point cloud is combined with a further set of points P_t collected with an optical tactile sensor, on which a *second* set of Gaussians G_t is initialised on tactile data. Details on the collection of the tactile data are reported in the Supplementary material. Specifically, each point in P_t serves as the mean for one of the Gaussians in G_t, while their color and scale attributes are initialised randomly. A uniform opacity value is assigned across all Gaussians in both G_t and G_c. Both sets of Gaussians can be cloned, split, and removed according to the heuristics outlined by Kerbl et al. [12]. The newly generated Gaussians that result from this process are assigned to the same set of their originals.

The optimisation of these Gaussians involves minimising the photometric loss between predicted and ground truth RGB image. This loss is further regularised by the edge-aware smoothness loss described in 4.2.2. In addition to the photometric and edge-aware smoothness loss, Gaussians G_t are regularised by the 3D transmittance loss, which we discuss in Section 4.2.1.

4.2. Regularisation

4.2.1 3D transmittance

The transmittance quantifies the degree to which light penetrates through a medium. Given N Gaussians and a point $\mathbf{p} \in \mathbb{R}^3$, where $\mathbf{p}$ belongs to the touch surface, we define the average 3D transmittance at $\mathbf{p}$ as:

$$\hat{T}(\mathbf{p}) = \frac{1}{N} \sum_{j=1}^{N} (1 - f(\mathbf{p}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\ \alpha_j), \tag{6}$$

where $\boldsymbol{\mu}_i \in \mathbb{R}^3$ and $\boldsymbol{\Sigma_i} \in \mathbb{R}^3$ are the mean and covariance of the unprojected 3D Gaussians. To manage the computational load and prioritise the optimisation of the Gaussians around the touch locations, we limit the number of Gaussians considered per point to those with the highest spatial influence on p. Moreover, we incorporate a distance-based filtering criterion designed to exclude Gaussians beyond a certain threshold distance from $\mathbf{p}$. This is because, in the process of selecting Gaussians based on their spatial impact, it is possible to include Gaussians that are positioned at considerable distances from the target point. Given that the Gaussian rasteriser subdivides the frustum into discrete tiles and calculates a point's transmittance solely from Gaussians within the same tile, distant Gaussians — even those with potential impact — would not account for the final transmittance at $\mathbf{p}$. We define our loss $\mathcal{L}_T$ as the average over all the $\mathbf{p} \in P_t$ points collected on the object surfaces:

$$\mathcal{L}_T = -\frac{1}{P} \sum_{i=1}^{P} \hat{T}(\mathbf{p_i}) \tag{7}$$

Minimising the average transmittance around the touch location effectively brings the Gaussians closer to the touched surfaces and increases their opacity, thereby implicitly forcing them to model the real underlying surface.

4.2.2 Edge-Aware Smoothness with Proximity-Based Masking

Given the sparsity of touch readings, we leverage an edge-aware smoothness loss $\mathcal{L}_\text{S}$ [8] to refine the surface reconstruction. This loss leverages the intuition that depth discontinuities often align with intensity changes in RGB images:

Figure 2. Proximity-Based Mask's impact on the gradients computed by the edge-aware smoothness loss. *Left*: points collected on the object's surface alongside the derived proximity mask. *Centre*: averaged horizontal and vertical gradients as determined by the smoothness loss, where lighter shades correspond to higher gradients. *Right*: integration of the smoothness loss with the proximity mask using two distinct approaches: **(a)** implementation of a distance-based Gaussian decay within the proximity mask, and **(b)** masking based on a discrete threshold from the contact surface.

$$\mathcal{L}_S = \frac{1}{N}\sum_{i,j}\left(|\partial_x D_{i,j}|e^{-\beta|\partial_x I_{i,j}|} + |\partial_y D_{i,j}|e^{-\beta|\partial_y I_{i,j}|}\right) \tag{8}$$

where D is the rasterised depth map and I is the ground truth image for a given view. The horizontal and vertical gradients are calculated by applying convolution operations with 5x5 Sobel kernels [29]. In contrast to the smaller kernel size commonly adopted by edge detectors, our filters capture more contextual information, reducing instability around specular highlights. While the edge-aware smoothness loss positively contributes to the refinement of surface reconstruction, its integration with the transmittance loss presents challenges. Specifically, the smoothness loss conflicts with the transmittance loss near touch locations — areas where we prioritise accurate contact surface reconstruction. Since $\mathcal{L}_S$ operates globally, it may suppress localised depth variations by reducing opacity, counteracting the transmittance loss.

To address this, we introduce a *Proximity-Based Masking* strategy, which modulates $\mathcal{L}_S$ based on distance to touch points. The creation of a proximity-based mask is achieved by calculating the 2D Euclidean distance transform [5] between every pixel in the rasterised depth map and its nearest projected 3D point. Then, a Gaussian decay is applied based on the calculated distance, producing the actual proximity mask. Finally, the edge-aware smoothness loss, which aims to align the gradients of the predicted depth map with those of the corresponding RGB image, is modulated by the inverse of this proximity mask, effectively diminishing the loss's influence in areas close to the contact surfaces and preserving edge details as the distance from touch locations increases. This integration balances local accuracy and global refinement: the transmittance loss ensures precise contact surfaces, while the edge-aware smoothness improves reconstruction in underexplored regions without interfering near touch constraints.

5. Results

5.1. Datasets & Evaluation

We train and evaluate our method on two object-centric datasets containing glossy and reflective surface: Shiny Blender [30] and Glossy Reflective [17]. The Shiny Blender [30] dataset consists of 6 objects characterised by non-Lambertian materials. For each object, the dataset provides 100 training images and 200 test images all accompanied by camera poses. The *ball* object was excluded in our evaluation due to the absence of ground truth depth maps, which makes the assessment of the reconstructed geometry infeasible. Although ground truth depth maps and normal maps are available, our model relies solely on RGB images and local depth maps collected on the object surface by simulating a touch sensor, and we never leverage ground truth depth maps for supervision.

The Glossy Synthetic [17] dataset consist of 8 objects, each represented by 128 images at a resolution of 800x800. The dataset is divided into 32 images for training and 96 images for testing. This dataset ranges from objects with large smooth surfaces displaying specular effects (e.g. *Bell, Cat*) to those with complex geometries (e.g. *Angel, Luyu*).

The evaluation metric employed to asses geometry reconstruction is the Chamfer Distance (CD) [25], which measures the similarity between predicted and ground truth point clouds. The Glossy Synthetic dataset provides the ground truth point cloud for CD calculation. For the Shiny Blender dataset, we created ground truth point clouds for CD calculation by projecting the ground truth depth maps in the three-dimensional space. This method ensures that we only consider areas visible to the camera, thereby accounting for occlusions and ensuring a fair comparison. The decision against using point clouds sampled directly from

Glossy Synthetic [17]				
Method	100 views		5 views	
	CD(↓)	SSIM(↑)	CD(↓)	SSIM(↑)
NeRO	0.0042	0.904	0.0586	0.779
3DGS	0.0075	0.933	0.0111	0.822
Ours[5 grasps]	0.0034	0.933	0.0026	0.828

Table 1. Average Chamfer Distance (CD) and SSIM on the Glossy Synthetic dataset [17]. Our method excels at recovering object geometry from glossy surfaces, outperforming baselines in both the 100-view and 5-view scenarios.

Shiny Blender[30]		
Method	CD(↓)	SSIM(↑)
3DGS	0.0037	0.948
3DGS + S	0.0028	0.950
3DGS + T[5 grasps]	0.0022	0.948
Ours[5 grasps]	0.0013	0.950

Table 2. Ablation on the full Shiny Blender dataset [30] (100 views). Our method, which considers both tactile data and our proposed smoothness loss, considerably improves the geometry reconstruction, while achieving comparable levels of image fidelity.

the ground truth meshes is aimed at avoiding discrepancies caused by occluded regions which are not comparable across different methods. For the predicted point clouds, we apply the method adopted on the Glossy Synthetic dataset. In addition, we provide a qualitative comparison between standard 3DGS and our proposed method on a dataset collected in the real-world (Sect. 5.4).

5.2. Glossy Synthetic

Full-views: Table 1 benchmarks our method against 3D Gaussian Splatting (3DGS) and NeRO. We chose these methods as they represent the state-of-the-art in terms of geometry reconstruction for challenging objects. Our method uses 5 grasps per object, whereas a grasp consists of "five" fingers touching the object, resulting in a total of 25 tactile readings. The results are reported for both the settings where all methods have full access to the 100 training views, and when they only have access to a minimal subset of 5 views. This section focuses on the 100-views scenario. To obtain the results of 3D Gaussian Splatting we trained using the official 3DGS implementation, while for NeRO we directly report the available results from [17] and values provided by the original authors. To compute the CD for 3D Gaussian Splatting-related methods, the predicted point cloud is derived from projecting rasterised depth maps to 3D, using the 96 test cameras. This results in an average of 500K points per object.

Our analysis demonstrates that our method significantly outperforms the leading competitor, 3DGS, in 3D reconstruction quality (see Table 1). With just five grasps, our approach achieves a 50% decrease in the average Chamfer Distance (CD), resulting in 0.0034 compared to 3DGS's 0.0075. This performance places our 3DGS-based method on par with, but in some cases ahead of, current NeRF-based state-of-the-art methods such as NeRO. Moreover, radiance field approaches such as NeRO are considerably more computationally demanding. For reference, while NeRO requires 25 hours of training [17], our method takes 1 hour on average to reconstruct an object with 5 grasps.

Minimal-views: We explore the performance of the different methods in a "minimal views" regime where models are provided with only 5 input views (Table 1 right columns). In this setting we evaluate the same methods considered in the previous experiment: 3DGS, NeRO and Ours. We used NeRO's official implementation to retrain the model. After training, we computed the CD for NeRO models as described in the original paper and extracted meshes using the Marching Cubes algorithm [18]. To ensure consistency in evaluation, we sampled around 500,000 points on the generated meshes, aligning with the previously detailed evaluation procedure for 3DGS. Figure 3 shows that in this setting the addition of grasps significantly improves the quality of the reconstructed geometry and multiview rendering. A comprehensive visualisation with all the objects in the Glossy Synthetic dataset is included in the Supplementary materials. Table 1 (right columns) reports quantitative results and shows how NeRO and 3DGS quality of 3D reconstruction drops dramatically when trained with few views, while our method can retain and, for some models, even improve. In this setting, our method achieves over 4 times improvement over the second best model in terms of geometry accuracy (0.0026 Ours vs 0.0111 3DGS), clearly showing the effectiveness of our formulation for high quality 3D reconstruction. The improved reconstruction contributes to enhanced quality in novel view synthesis, as demonstrated by the SSIM scores.

5.3. Shiny Blender

We continue the evaluation on the Shiny Blender dataset [30] where we focus on methods based on 3D Gaussian Splatting. We are interested in exploring the impact of the different design choices and as such we compare our method against a baseline defined by vanilla 3DGS, a method that only uses the edge-aware smoothness regularisation introduced in Section 4.2.2 (3DGS+S), and a method that only uses 3DGS and tactile information (3DGS + T).

Table 2 reports quantitative results comparing the four variants of Gaussian Splatting-based methods according to both the quality of the 3D reconstruction (CD) and of novel view synthesis (SSIM). Our method does not sacrifice the

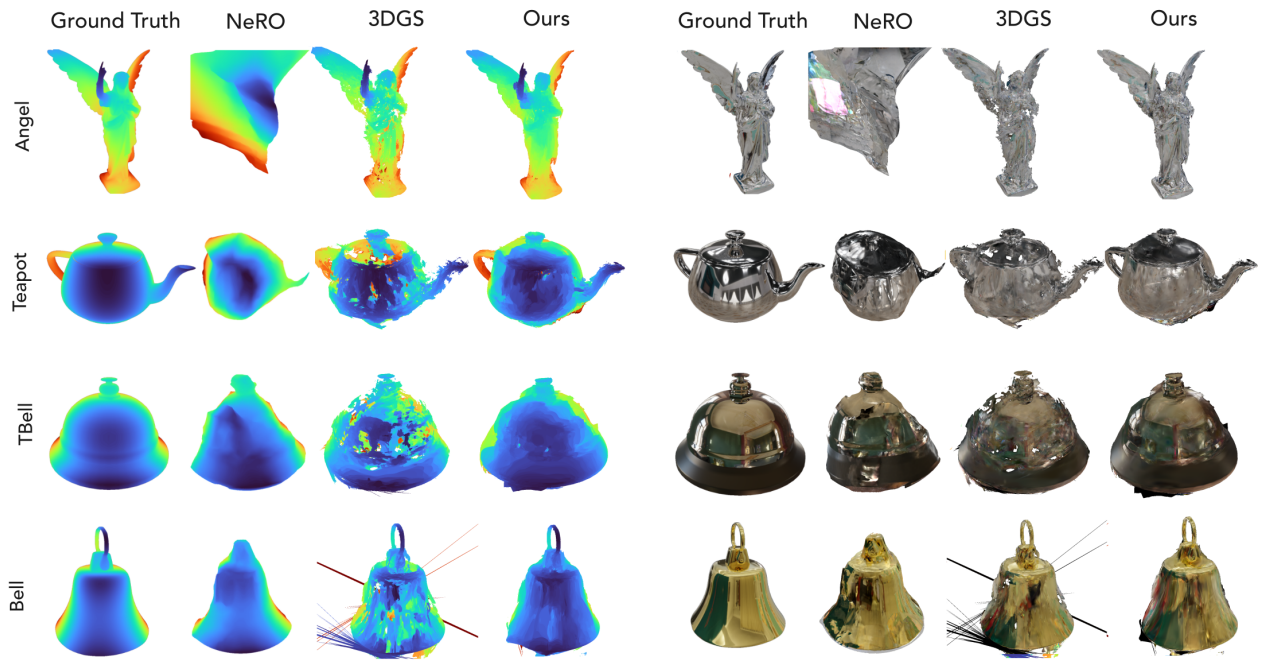

Figure 3. Surface reconstruction and novel view synthesis qualitative results using 5 training views on a sample of the Glossy Synthetic dataset.

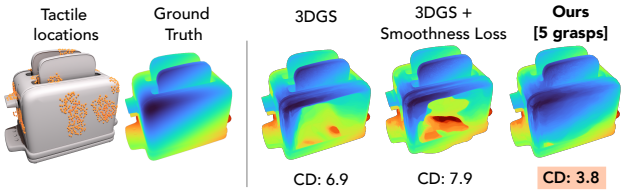

Figure 4. Rasterised depth maps of the *Toaster* object. Our method results in a smoother, accurate reconstruction compared to 3DGS and 3DGS with additional regularisation on the smoothness loss.

Method	Toaster	Car	Coffee	Helmet	Teapot	Avg.
3DGS	0.53	0.69	0.64	0.61	0.79	0.65
3DGS + S	0.53	0.70	0.60	0.67	0.79	0.66
Ours[5 grasps]	0.60	0.74	0.64	0.70	0.79	0.70

Table 3. Comparison of 3D reconstruction methods on the full Shiny Blender dataset (100 views). We report the Area Under the Curve (AUC) for prediction accuracy at increasing distance thresholds between predicted and ground truth point clouds.

quality of the rendering as shown by neutral metrics in terms of SSIM compared to 3DGS, while it significantly improves the geometry of the reconstructions as measured by average CD (1.3 Ours vs 3.7 3DGS). The gap is more pronounced on models with relatively large shiny surfaces like *toaster* and *helmet*. A qualitative visualisation of the difference in reconstructions for the *toaster* object is shown in Figure 4. The results of 3DGS+S also show how the smoothness loss is effective to improve the quality of 3D reconstruction, but it still falls short compared to our full method where constraints given by the surfaces reconstructed by the grasps and the smoothness regularisation act in synergy to further boost performance. Similarly, while the addition of touches in 3DGS+T results in better reconstruction quality over the standalone 3DGS model, the absence of smoothness regularisation limits its ability to refine the reconstruction beyond the areas of contact.

We observed that the Chamfer Distance sometimes leads to inconsistencies between quantitative results and visual quality, as it may overlook structural discrepancies such as gaps and holes in the predicted point cloud. This occurs because CD considers only the minimum distance from each point in the predicted point cloud to the nearest point in the ground truth, and vice versa. To address this limitation, we implemented a Point Cloud Coverage metric as the completeness of the predicted point cloud relative to the ground truth. It involves verifying the presence of at least one predicted point within progressively increasing radii around each ground truth point and plotting these results to form a coverage curve (Figure 5). A comprehensive analysis across all tested objects is shown in the Supplementary material. In Table 3 we report the area under the curve (AUC) which quantify the overall accuracy of the reconstruction, whereas higher values indicating more comprehensive coverage of the ground truth geometry. Our results indicate that com-

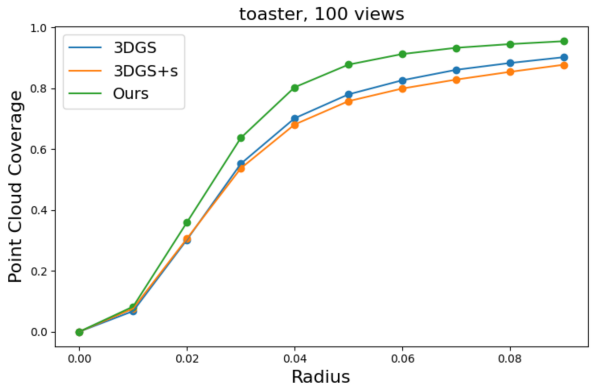

Figure 5. Point Cloud Coverage curves for the *Toaster* object. Combining tactile and visual data leads to higher accuracy at lower distance thresholds.

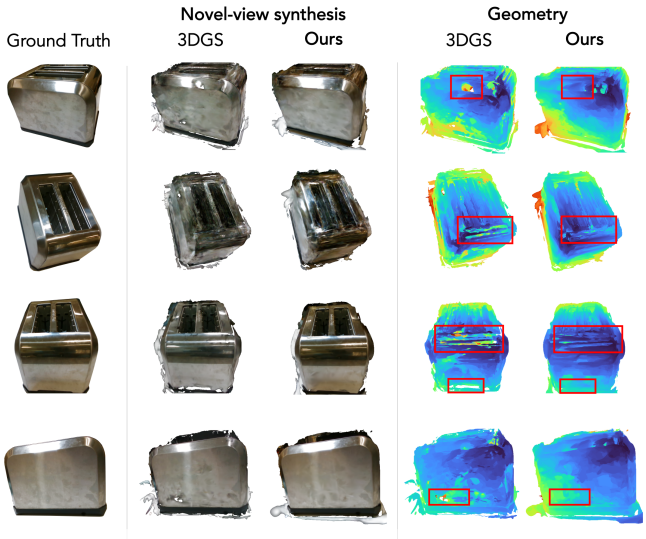

Figure 6. Qualitative reconstructions in the real-world. Highlighted areas show that the introduction of touches leads to smoother and more accurate surface reconstruction. SSIM on the generated views in this figure: 0.78 (3DGS) vs. 0.82 (Ours) – higher is better.

bining visual with contact data achieves higher accuracy at lower radius thresholds and therefore a better reconstruction quality.

5.4. Real world experiment

In addition to the experiments in simulated environments, we conduct a qualitative comparison on real world reconstruction. We employed a robotic hand (Allegro Hand) equipped with four tactile sensors and an in-hand camera to acquire multiple views and grasps of a real object, aiming to reconstruct its geometry using both a standard 3D Gaussian Splatting model (3DGS) and our novel approach. Specifically, we collected 25 views and 20 tactile readings to capture the object's geometry. The qualitative results presented here (see Figure 6) are part of an evaluation to demonstrate the improvements our method offers in real-world geometric reconstruction of challenging surfaces. Due to the unavailability of ground truth geometry, the calculation of the Chamfer Distance is not feasibile. For clearer visualisation and to emphasise the quality of object reconstruction, we presented cropped images and 3D Gaussian representations focused on the object, though training utilised the original, uncropped images.

Our results highlight regions where our approach produces smoother surfaces or eliminates holes as well as generating more accurate novel view synthesis. These findings align with outcomes from simulated experiments, indicating that the inclusion of local depth maps obtained from tactile sensors improves reconstruction quality. For more information on the data collection procedure employed in this experiment, please refer to the supplementary material.

6. Limitations and Conclusion

This work introduces a novel approach for incorporating touch sensing with vision. We believe to be the first to have explored the problem of reconstruction and novel view synthesis of an object that is both seen and touched. We presented convincing results both qualitatively and quantitatively on two complex datasets, and we showed that our method outperforms baselines, especially when a set of minimal views of the object are used.

Despite these promising results, there are some limitations that highlight areas for future research. The current methodology employs random touch sampling to collect contact surfaces for scene reconstruction. This strategy may not always be efficient and could be extended by an adaptive sampling method, which can select sampling locations to complement visual data. Future work could focus on investigating how multimodal interaction can further improve reconstruction of transparent objects within the scene. We believe incorporating ideas like surface modelling and surface normals can also improve upon our results.

Acknowledgements

We would like to thank Alex Zook and Stan Birchfield for their feedback regarding the scope and contributions of our research, which helped improve the quality of our work.

References

[1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF Inter-*

national Conference on Computer Vision, pages 5855–5864, 2021. 3

[2] Alex Church, John Lloyd, Nathan F Lepora, and others. Tactile sim-to-real policy transfer via real-to-sim image translation. In *Conference on Robot Learning*, pages 1645–1654. PMLR. 1

[3] Mauro Comi, Yijiong Lin, Alex Church, Alessio Tonioni, Laurence Aitchison, and Nathan F Lepora. Touchsdf: A deepsdf approach for 3d shape reconstruction using vision-based tactile sensing. *IEEE Robotics and Automation Letters*, 2024. 3, 1, 2

[4] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. http://pybullet.org, 2016–2021. 1

[5] Ricardo Fabbri, Luciano Da F Costa, Julio C Torelli, and Odemir M Bruno. 2d euclidean distance transform algorithms: A comparative survey. *ACM Computing Surveys (CSUR)*, 40(1):1–44, 2008. 5

[6] Ruohan Gao, Yiming Dou, Hao Li, Tanmay Agarwal, Jeannette Bohg, Yunzhu Li, Li Fei-Fei, and Jiajun Wu. The objectfolder benchmark: Multisensory learning with neural and real objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17276–17286, 2023. 2

[7] Yuan-Chen Guo, Di Kang, Linchao Bao, Yu He, and Song-Hai Zhang. Nerfren: Neural radiance fields with reflections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18409–18418, 2022. 3

[8] Philipp Heise, Sebastian Klose, Brian Jensen, and Alois Knoll. Pm-huber: Patchmatch with huber regularization for stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2360–2367, 2013. 4

[9] Muhammad Zubair Irshad, Mauro Comi, Yen-Chen Lin, Nick Heppert, Abhinav Valada, Rares Ambrus, Zsolt Kira, and Jonathan Tremblay. Neural fields in robotics: A survey. *arXiv preprint arXiv:2410.20220*, 2024. 3

[10] Yingwenqi Jiang, Jiadong Tu, Yuan Liu, Xifeng Gao, Xiaoxiao Long, Wenping Wang, and Yuexin Ma. Gaussianshader: 3d gaussian splatting with shading functions for reflective surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5322–5332, 2024. 3

[11] Brian Karis and Epic Games. Real shading in unreal engine 4. *Proc. Physically Based Shading Theory Practice*, 4(3):1, 2013. 3

[12] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 2, 3, 4

[13] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, et al. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020. 2, 1

[14] Nathan F Lepora. Soft biomimetic optical tactile sensing with the tactip: A review. *IEEE Sensors Journal*, 21(19): 21131–21143, 2021. 2, 1

[15] Yijiong Lin, John Lloyd, Alex Church, and Nathan F Lepora. Tactile gym 2.0: Sim-to-real deep reinforcement learning for comparing low-cost high-resolution robot touch. *IEEE Robotics and Automation Letters*, 7(4):10754–10761, 2022. 1

[16] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 2

[17] Yuan Liu, Peng Wang, Cheng Lin, Xiaoxiao Long, Jiepeng Wang, Lingjie Liu, Taku Komura, and Wenping Wang. Nero: Neural geometry and brdf reconstruction of reflective objects from multiview images. In *SIGGRAPH*, 2023. 2, 3, 5, 6

[18] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998. 6

[19] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3

[20] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 3

[21] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 3

[22] Haozhi Qi, Brent Yi, Sudharshan Suresh, Mike Lambeta, Yi Ma, Roberto Calandra, and Jitendra Malik. General in-hand object rotation with vision and touch. In *Conference on Robot Learning*, pages 2549–2564. PMLR, 2023. 3

[23] Edward Smith, Roberto Calandra, Adriana Romero, Georgia Gkioxari, David Meger, Jitendra Malik, and Michal Drozdzal. 3d shape reconstruction from vision and touch. *Advances in Neural Information Processing Systems*, 33: 14193–14206, 2020. 3, 1

[24] Edward Smith, David Meger, Luis Pineda, Roberto Calandra, Jitendra Malik, Adriana Romero Soriano, and Michal Drozdzal. Active 3d shape reconstruction from vision and touch. *Advances in Neural Information Processing Systems*, 34:16064–16078, 2021. 3, 1

[25] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2974–2983, 2018. 5

[26] S. Suresh, Z. Si, J. Mangelson, W. Yuan, and M. Kaess. ShapeMap 3-D: Efficient shape mapping through dense

touch and vision. In *Proc. IEEE Intl. Conf. on Robotics and Automation, ICRA*, Philadelphia, PA, USA, 2022. 2

[27] Aiden Swann, Matthew Strong, Won Kyung Do, Gadiel Sznaier Camps, Mac Schwager, and Monroe Kennedy III. Touch-gs: Visual-tactile supervised 3d gaussian splatting. *arXiv preprint arXiv:2403.09875*, 2024. 3

[28] Michael W Tao, Jong-Chyi Su, Ting-Chun Wang, Jitendra Malik, and Ravi Ramamoorthi. Depth estimation and specular removal for glossy surfaces using point and line consistency with light-field cameras. *IEEE transactions on pattern analysis and machine intelligence*, 38(6):1155–1169, 2015. 2

[29] Manoj K Vairalkar and SU Nimbhorkar. Edge detection of images using sobel operator. *International Journal of Emerging Technology and Advanced Engineering*, 2(1):291–293, 2012. 5

[30] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022. 2, 3, 5, 6

[31] Shaoxiong Wang, Jiajun Wu, Xingyuan Sun, Wenzhen Yuan, William T. Freeman, Joshua B. Tenenbaum, and Edward H. Adelson. 3d shape perception from monocular vision, touch, and shape priors. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1606–1613. IEEE. 2

[32] Max Yang, Chenghua Lu, Alex Church, Yijiong Lin, Chris Ford, Haoran Li, Efi Psomopoulou, David AW Barton, and Nathan F Lepora. Anyrotate: Gravity-invariant in-hand object rotation with sim-to-real touch. *arXiv preprint arXiv:2405.07391*, 2024. 2

[33] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017. 2, 1