2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Automatic Physically-Based Sim2Real for Tactile Images through Differentiable Path-Tracing Rendering*

Guillaume Duret[1,3], Anna Samsonenko[1], Florence Zara[2], Jan Peters[3] and Liming Chen[1]

Abstract— **High-fidelity simulation of vision-based tactile sensors is essential for developing data-driven robotic manipulation algorithms. However, a significant sim-to-real gap persists due to the difficulty in modeling complex optical effects, such as refraction through protective glass layers, and in accurately estimating physical parameters like sensor pose and lighting. To bridge this gap, we introduce a novel, fully differentiable pipeline for visual tactile simulation. Leveraging a differentiable path tracer, our method optimizes critical parameters—including camera pose, lighting conditions, and object texture—directly from just three real images. This approach achieves highly realistic simulations with physically accurate light transport and glass refraction. We validate our method through a comprehensive benchmark against real-world data, demonstrating state-of-the-art sim-to-real accuracy. We also enable novel applications, such as mesh reconstruction from a single tactile image via inverse rendering. To overcome the computational cost of path tracing, we further use a image-to-image translation model. This model uses high-fidelity simulated data alongside Normalized Object Coordinate Space (NOCS) maps as input, preserving crucial deformation information while enabling rapid inference. The code is available on `https://tacdiffrend.github.io/`**

I. INTRODUCTION

Vision-based tactile sensors have emerged as a powerful tool for providing rich, high-resolution contact feedback in robotic manipulation [13]. Their effectiveness in tasks ranging from slip detection to object recognition and reinforcement learning is heavily dependent on large, high-quality datasets. Physical simulation offers a scalable and cost-effective alternative to cumbersome real-world data collection.

Early simulation methods relied on rigid-body dynamics with depth cameras and Gaussian smoothing to approximate gel deformation [20], [16]. While efficient, these approaches usually fail to capture the shear forces and marker displacements that are a primary source of tactile information. Recent work has therefore shifted to deformable simulation methods like the Finite Element Method (FEM) [7], [22], which provide higher fidelity by modeling the sensor's soft body and tracking embedded markers.

Despite these advances, a significant visual sim-to-real gap remains. A primary challenge is the complex light transport within the sensor, particularly refraction through its protective glass layer, which distorts the apparent position and shape of markers. Previous rendering techniques have used simplified rasterization with manual LED effects [20], ray tracing for shadows [1], or post-hoc corrections based on Snell's law [22]. These methods often require manual parameter tuning and assume known camera poses, which can vary between sensor units due to manufacturing tolerances.

In this work, we address these limitations by introducing a fully differentiable rendering pipeline for visual tactile sensors. Our core contribution is a method to automate the simulation setup through the precise, automatic optimization of critical parameters—including camera pose, lighting, and texture—directly from minimal real-world data. This is achieved by leveraging a differentiable path tracer (Mitsuba 3 [11]) to compute gradients through the entire physical rendering process, including refraction. Furthermore, to enable near real-time application, we train a fast Pix2Pix image-to-image translation model that converts expressive Normalized Object Coordinate Space (NOCS) map representations into photorealistic tactile images.

Our key contributions are:

- **The first fully differentiable rendering pipeline for visual tactile sensors**, which enables gradient-based optimization of physical parameters (pose, texture, lighting) directly from real images.
- **A novel optimization method for visual tactile simulation** that explicitly models key optical effects like glass refraction and RGB lighting, significantly narrowing the sim-to-real gap.
- **A framework for generating high-fidelity synthetic data** via image-to-image translation from NOCS maps, achieving state-of-the-art visual fidelity and enabling fast, near real-time inference for downstream tasks.
- **Extensive validation on a real robotic platform**, demonstrating effective sim-to-real transfer and novel applications in inverse problems such as tactile geometry reconstruction.

II. RELATED WORK

A. Physical Simulation

Pioneering simulations of visual tactile sensors initially focused on rigid-body dynamics with interpenetration [12],

*This work was in part supported by the French Research Agency, l'Agence Nationale de Recherche (ANR), through the projects Learn Real (ANR-18-CHR3-0002-01), Chiron (ANR-20-IADJ-0001-01), Aristotle (ANR-21-FAI1-0009-01), Astérix (ANR-23-EDIA-0002), Demeter (ANR-25-HTCE-0002) and Protheus (ANR-25), the French national investment prioritary program PSPC FAIR WASTE project, as well as a donation to Fonds de Dotation Centrale Lyon by Huawei Technologies RD France. It was granted access to the HPC resources of IDRIS under the allocation 2025-[AD011015271R1], 2025-[AD011015591R1] and 2026-[A0191013894] made by GENCI.

[1]Centrale Lyon, CNRS, LIRIS, UMR5205, F-69130 Ecully, France, guillaume.duret@ec-lyon.fr

[2]UCBL, CNRS, LIRIS, UMR5205, F-69622 Villeurbanne, France

[3]Intelligent Autonomous Systems Lab, Technical University of Darmstadt, 64289 Darmstadt, Germany

[20]. In these models, both the object and the tactile sensor were treated as rigid bodies. Tactile images were generated using a simulated depth camera. Subsequent work added Gaussian post-processing to the depth-based data to mimic gel deformation and avoid excessively sharp contact geometries [8], [16]. This method has been widely used to generate tactile data for classification tasks and, more recently, for reinforcement learning due to its computational efficiency [14], [2]. However, rigid-body simulation primarily captures vertical forces and deformation, failing to accurately model translational and rotational shear forces. These shear forces are critical as they drive marker displacement within the gel, which is a major source of information in the resulting tactile image. To achieve higher fidelity, deformable simulation methods such as the Material Point Method (MPM) [5] and the Finite Element Method (FEM) [7], [6] have been adopted. These approaches enable marker-based simulations of tactile sensors, where markers are treated as a direct texture on the deformable mesh, allowing them to move with the mesh. In this work, we build upon the FEM-based simulation from TacFlex [22], which has been shown to offer a favorable trade-off between speed, accuracy, and sim-to-real transfer. We extend this foundation by developing a novel differentiable rendering component.

B. Visual Simulation

The rendering techniques for visual tactile sensors have evolved significantly. Initial approaches used basic rasterization with manually positioned LED effects [8], later improved with added backgrounds [20] and calibration techniques [20]. Ray tracing was subsequently introduced to produce more realistic shadow effects and higher-quality images [1]. With the advent of deformable simulations, marker-based rendering was able to model the displacement of markers due to shear, translation, and rotation by linking marker positions to the mesh itself as a texture. Early methods used the undeformed tactile image as a direct texture [7], while others rendered only the marker positions [3]. TacFlex [22] combined markers as textures with a background image and applied a manual glass effect correction based on Snell's law for optical refraction. In this work, we adopt a similar structure but leverage a fully differentiable pipeline to automatically optimize all rendering parameters. Key differences include integrating the glass effect directly into a physically-based light transport model instead of applying a separate correction. Furthermore, rather than using a complete background image, we simulate the lighting on a uniform-color mesh to match the real sensor's construction, and we optimize the camera pose, which was previously assumed to be known.

An alternative approach in rendering is data-driven image-to-image translation methods like Pix2Pix [12], which offer fast inference and high visual quality but require thousands of paired real and synthetic images for training. While previous works have attempted to reduce the cost of data annotation by using CycleGAN [4], [23], [15] to learn from unpaired datasets, our approach benefits from a key simplification: we use high-fidelity simulated data as a direct proxy for real data, ensuring perfect pixel alignment and justifying the use of Pix2Pix [10]. This allows for more direct supervision during training. Although diffusion models [9] have been explored for tactile sensors, their significantly slower inference speed makes them less practical. Since tasks for visual tactile sensors often perform well with simpler models, we employ a Pix2Pix architecture to achieve state-of-the-art inference speed.

C. Differentiable Simulation for Tactile Sensing

Differentiable simulations are crucial for enabling efficient gradient-based learning for various tasks, often proving more sample-efficient than reinforcement learning methods [17]. This research direction is emerging for visual tactile sensors but has remained largely focused on differentiating through marker displacement [21]. For instance, [17] demonstrated the first soft-body-based differentiable tactile simulation using Material point methods (MPM), though their differentiable pipeline was limited to marker positions for manipulation tasks. In our work, we focus on the differentiable generation of the final tactile image itself. We employ Mitsuba 3.0 [11], a differentiable path-tracing renderer, to bridge the gap between deformable simulation and photorealistic output. To the best of our knowledge, this constitutes the first use of full differentiable rendering for visual tactile sensors. This approach allows for the optimization of simulation and rendering parameters (e.g., material properties, lighting, camera pose) from minimal real-world data. It enables physically accurate modeling of effects like subsurface scattering and specular reflections from the protective glass. Furthermore, the differentiability enables solving inverse problems, such as reconstructing the deformed mesh geometry from an observed tactile image.

III. Differentiable rendering for visual tactile sensors

A. Differentiable rendering advantages

The differentiable rendering pipeline of Mitsuba 3.0 [11] is particularly well-suited for visual tactile simulation due to the unique characteristics of the domain. Unlike general computer vision applications where complex geometry and severe occlusions create challenging discontinuities for gradient-based optimization, tactile images feature smooth deformable surfaces and continuous contact areas, ensuring stable gradient computation. Mitsuba's key advantage lies in its ability to physically model three crucial aspects: 1) light transport through the protective glass layer via accurate BSDF modeling, including refraction to model the zoom effect, 2) natural shadow formation through path tracing that captures global illumination effects, and 3) differentiable parameter optimization through gradients. The differentiability is achieved through the gradient of the rendering equation:

$$\frac{\partial I}{\partial \theta} = \frac{\partial}{\partial \theta} \int_{\Omega} L_i(\omega_i), f_r(\omega_i, \omega_o), (\mathbf{n} \cdot \omega_i) d\omega_i \qquad (1)$$

where θ represents any render parameter (material properties, geometry, lighting), enabling efficient optimization of simulation parameters to minimize the difference between synthetic and real tactile images through gradient descent methods. Figure1 illustrates the overall parameter-optimization pipeline step by step: first, pose optimization (SectionIII-B), including glass refraction; then mesh uniform-color adjustment, lighting-intensity and pose correction, and background-light refinement (SectionIII-C); and finally, texture optimization to add markers (SectionIII-D).

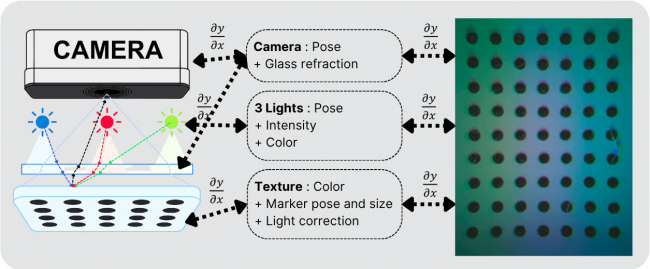


Fig. 1: Illustration of the overall pipeline for optimizing scene parameters from a real tactile image, including glass effects, camera pose, texture and marker configuration, as well as lighting color, pose, and intensity.

B. Pose Optimization

1) Initial Stage: Optimization Without Glass: The initial stage of our optimization pipeline addresses pose estimation, a critical challenge in visual tactile sensing due to manufacturing inconsistencies across sensor instances. These variations, present even in commercial sensors like the GelSight Mini, introduce subtle deviations from perfect alignment that significantly impact precision manipulation tasks.

Our approach minimizes data requirements by utilizing only three inputs: a single real tactile image, a mesh model of the tactile sensor, and the known marker grid specifications (9×7 array with 2mm×2.05mm spacing and 1mm diameter markers). From these specifications, we generate a perfect binary texture representing the marker pattern, with the textured mesh initially placed at a random pose for optimization.

The optimization process leverages circular markers visible in the reference image through the loss function:

$$\mathcal{L}_{\text{pose}} = \sum_{j=1}^{N} \left((u_j - u_j^{\text{ref}})^2 + (v_j - v_j^{\text{ref}})^2 \right) \tag{2}$$

where u_j, v_j represent projected marker coordinates and u_j^{ref}, v_j^{ref} correspond to positions detected via circle Hough transform. To maintain differentiability, we utilize 3D marker positions in the mesh's local coordinate system with a differentiable 3D-to-2D projection.

2) Refinement Stage: Optimization With Glass: The introduction of glass refraction necessitates significant methodological modifications to the pose optimization pipeline. While the initial stage employed a simple pinhole camera model for projection, this approach becomes inadequate when accounting for the complex optical refraction effects introduced by the protective glass layer, as illustrated in Fig 2.

To address this challenge while preserving differentiability, we implement a refined optimization approach using Mitsuba's differentiable rendering capabilities. The Stage 2 loss function employs a **masked Mean Squared Error (MSE) loss** that incorporates physical light transport simulation through glass refraction:

$$\mathcal{L}_{\text{stage2}} = \sum_{i=1}^{N} M_i \cdot (I_{\text{rendered},i} - I_{\text{ref},i})^2 \tag{3}$$

where M_i represents the binary mask value at pixel i targeting marker regions, $I_{\text{rendered},i}$ is the rendered binary pixel value incorporating glass refraction effects, and $I_{\text{ref},i}$ is the reference binary pixel value. This formulation focuses the optimization specifically on critical marker regions while ignoring irrelevant background pixels.

This approach maintains the same optimization framework but computes marker positions through physically-based ray tracing rather than simple projection. The binary mask $\mathbf{M}$ is generated using circle detection on the reference image, ensuring only marker regions contribute to the gradient computation.

The refinement stage produces substantially improved marker alignment, eliminates scaling discrepancies introduced by glass refraction, and yields a physically realistic textured mesh that deforms identically to the real sensor while maintaining proper scaling relationships. This approach enables precise sensor pose estimation from minimal real-world data while accounting for complex optical effects through the protective glass layer, establishing a foundation for high-fidelity tactile simulation.

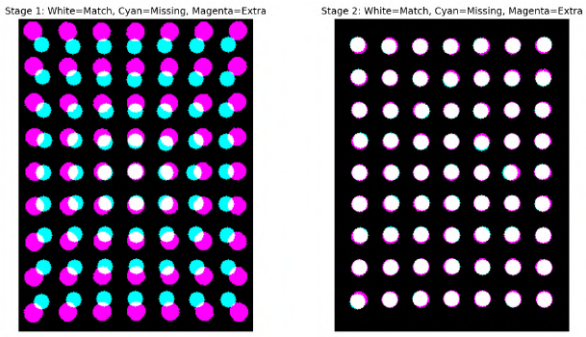


Fig. 2: Comparison of optimization results with and without glass refraction modeling. The left panel shows results from Stage 1 (without glass modeling), demonstrating noticeable misalignment of markers. The right panel shows Stage 2 results (with glass modeling), showing precise alignment of both position and scale. Color coding indicates: white (perfect matches), cyan (missing markers), and magenta (extra markers).

As demonstrated in Fig 2, the glass effect has a significant

impact on marker appearance. The validity of our approach is verified through three metrics: (1) marker alignment precision, (2) size consistency between simulated and real markers, and (3) measurement of marker patterns on textured meshes. These validations confirm that our method achieves perfect matching with real visual tactile sensors while maintaining size coherence between the mesh and markers.

C. Lighting optimization

1) Joint Lighting and Material Optimization: The next critical step in our pipeline is lighting optimization. While previous work often approximates lighting with background environments [20], [22], our differentiable rendering approach requires physically accurate light representations to ensure reliable simulation results. We initialize our scene with three rectangular area lights positioned around the sensor perimeter, matching the physical sensor design, along with a mesh featuring uniform coloration that approximates the real sensor's appearance. The optimization process simultaneously adjusts multiple photometric parameters:

- Light properties: color and intensity for each of the three light sources
- Material properties: diffuse reflectance (color) of the object mesh

We employ a dual-reference optimization strategy using two distinct target images with complementary purposes. First, a texture reference image captured without deformation provides global color information for material optimization. This reference allows us to recover the intrinsic mesh color while accounting for the collective illumination effect of all three light sources.

Second, a tactile image featuring a deformed cross pattern provides precise lighting information. We selected this pattern because it reveals lateral illumination characteristics from all three light directions, enabling accurate light parameter optimization that generalizes to arbitrary shapes and deformations.

The complete loss function combines weighted contributions from both reference targets:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{light}} + \beta \cdot \mathcal{L}_{\text{mesh}} \tag{4}$$

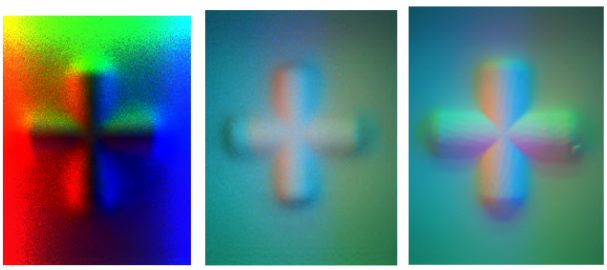
where α and β are weighting coefficients (typically 0.3 and 0.7 respectively), and the individual loss components are defined as:

$$\mathcal{L}\text{light} = \frac{1}{N}\sum_{i=1}^{N} W_i \cdot \left(T(I_{\text{rendered},i}) - T(I_{\text{ref-light},i})\right)^2 \tag{5}$$

$$\mathcal{L}\text{mesh} = \frac{1}{N}\sum_{i=1}^{N} \left(T(I_{\text{rendered},i}) - T(I_{\text{ref-mesh},i})\right)^2 \tag{6}$$

where $T(x) = x/(x + 1.0)$ applies a tone mapping operator, and W_i represents an adaptive weight map that emphasizes regions with higher reconstruction error and focus strong lighting area

This comprehensive optimization pipeline enables high-fidelity sim2real transfer for visual tactile sensors, accurately reproducing both visual appearance and physical light interaction behaviors as illustrated in Fig 1 and Fig 3.



(a) Initial Scene (b) Final Scene (c) GT Real

Fig. 3: Qualitative results of the lighting optimization. From left to right: the initial state of the scene, the final optimized result, and the ground-truth image.

D. Marker Texture Optimization

Given the accurately estimated mesh–camera pose, mesh color, and 3D light intensity and pose parameters, we focus on optimizing the texture representation specifically for the marker regions. This approach targets only the circular markers, which are the primary features of interest for tactile sensing. The optimization minimizes the Mean Squared Error (MSE) between the rendered marker patterns and the reference image:

$$\mathcal{L}_{\text{marker}} = \frac{1}{N_{\text{markers}}} \sum_{i \in \mathcal{M}} \left(I_{\text{rendered},i} - I_{\text{ref},i}\right)^2 \tag{7}$$

where $\mathcal{M}$ denotes the set of pixels belonging to marker regions, and N_{markers} is the total number of marker pixels. This targeted optimization ensures precise reproduction of marker appearance while maintaining computational efficiency. The final result is a texture with accurately embedded markers.

IV. NOCS-Based Image-to-Image Translation

Our proposed simulation method achieves high quality physical realism, accurately modeling complex optical effects including glass refraction, and ease the ability to optimize simulation parameters. However, this physical accuracy comes at a computational cost, as path tracing with differentiable rendering through refractive interfaces remains more computationally intensive than rasterization-based or learned approaches. To address this limitation while preserving the benefits of our high-fidelity simulation, we train a NOCS-based [18] image2image model [10] translation framework.

A. NOCS Representation for Tactile Sensing

Unlike conventional approaches that use simulated depth maps, our method leverages the comprehensive data generated by our physical simulation pipeline. We employ Normalized Object Coordinate Space (NOCS) maps [18], which provide a more expressive representation for tactile deformation. While depth maps primarily capture object

contact geometry, NOCS maps additionally encode marker displacement patterns resulting from translation and shear rotation—critical features for accurate tactile interpretation. These maps are rendered directly from our FEM-based simulations, providing ground truth correspondence between visual appearance and underlying deformation states.

B. Dataset Generation

Our dataset generation process illustrated in 4 employs 10 distinct indenters corresponding to those used in real-world experiments 5. For each indenter, we initiate 50 simulation trials with random initial positions and rotations above the sensor surface. Each simulation sequence includes: (1) a vertical compression phase, (2) four directional translations, and (3) a final rotational motion. To ensure dataset diversity and avoid duplication, we preserve only the forward simulation path, using the final state of each phase as the initial condition for subsequent motions. This approach prevents error accumulation in the FEM simulation while significantly accelerating data generation. The complete dataset comprises over 56,000 samples, each containing rendered RGB images, depth maps, surface meshes and corresponding NOCS representations.

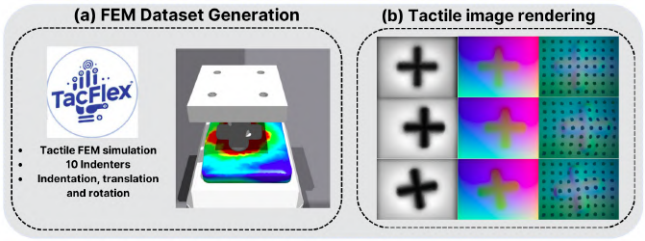


Fig. 4: Illustration of the dataset generation process using the FEM-based simulator TacFlex [22], extended to render Normalized Object Coordinate Space (NOCS) representations alongside RGB and Depth data. The system captures mesh geometry, RGB images, depth maps, and NOCS maps across three deformation scenarios: axial compression (first row), four-directional translation (second row), and dual-sensor rotation (third row).

V. Experiment and Results

A. Benchmarking on Indenter-based Quality Baseline

To evaluate our method comprehensively, we establish a rigorous benchmarking protocol using the indenter set shown in Figure 5. For fair comparison and to isolate rendering quality from other factors, we utilize FEM simulations that provide ground truth surface deformation data through physical integration. We optimize our simulation parameters using the pipeline described in Section III, ensuring identical reference images for both baseline methods and our approach.

Our ground truth dataset extends beyond conventional benchmarks by including not only vertical indentation (2 mm depth) and lateral translations of 17mm, but also rotational deformations of menus and plus 6°—a critical aspect often neglected in prior work that primarily focuses on vertical forces and depth maps. This comprehensive dataset enables evaluation under complex multi-axis loading conditions.

We validate our method through extensive quantitative analysis, demonstrating state-of-the-art performance in sim2real transfer accuracy. Our evaluation encompasses both image-based metrics (MSE, PSNR, SSIM, SMAPE) and geometry-aware measures including marker position accuracy and deformation field consistency. The results shows improvements over existing methods across all evaluation metrics, particularly in capturing shear-induced deformations and rotational artifacts.

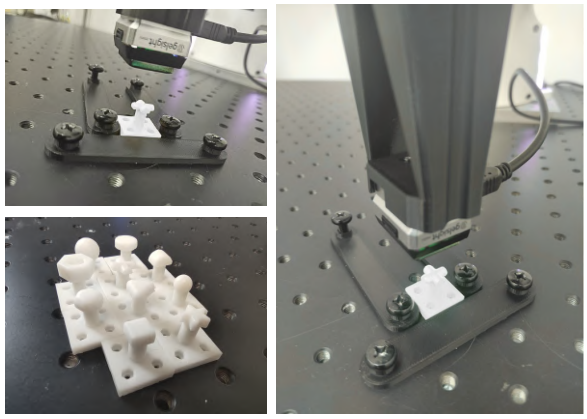


Fig. 5: Experimental setup for real-world tactile data acquisition using a GelSight Mini sensor mounted on a calibrated Franka Panda robotic arm. The data collection protocol involves: (a) indenting the sensor surface, (b) translating 2,mm in four cardinal directions, and (c) performing $\pm 5^{\circ}$ rotations to capture shear responses. This comprehensive benchmark encompasses 10 distinct indenters to evaluate performance across diverse contact geometries.

B. Mesh Optimization through Differentiability

This section demonstrates a direct application of our differentiable rendering pipeline for solving inverse problems in tactile sensing. Given a single real tactile image and our pre-optimized renderer, we leverage differentiability to reconstruct the underlying deformed geometry of the tactile sensor of the cross indenter. Unlike traditional approaches that render RGB images from known deformed meshes, our method operates in reverse: we optimize the mesh parameters to minimize the difference between the rendered and real tactile images.

Formally, we solve the inverse rendering problem:

$$\mathbf{V}^* = \arg\min_{\mathbf{V}} \mathcal{L}(I_{\text{render}}(\mathbf{V}), I_{\text{real}}) \tag{8}$$

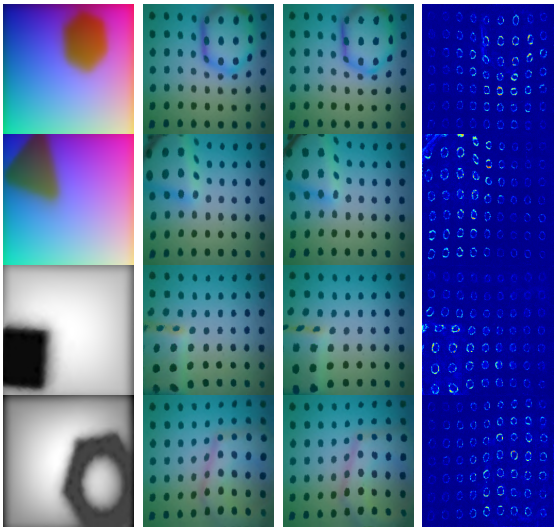
where $\mathbf{V}$ represents the mesh vertex positions, I_{render} is the image rendered using our differentiable pipeline, and I_{real} is the target real tactile image. The loss function $\mathcal{L}$ combines both photometric and geometric terms to ensure accurate reconstruction.

We evaluate our approach using the benchmark described in Section V-A, utilizing real tactile images as input and comparing the reconstructed meshes against ground truth FEM-

based deformations. The optimization proceeds through gradient descent, with gradients flowing backward through the entire rendering pipeline including light transport, refraction, and surface interactions.

This inverse optimization capability demonstrates the full potential of our differentiable framework, enabling not only realistic image synthesis but also accurate geometric optimisation from single tactile observations—a crucial capability for tactile-based manipulation and shape recognition tasks.

C. Result of Pix2pix model



(a) Real Sim (b) Fake Real (c) GT Real (d) Error

Fig. 6: Qualitative results of Pix2Pix inference. The top two rows show model outputs translating NOCS maps into RGB renderings, while the bottom two rows depict inference from depth images to RGB. Each triplet consists of the input (left), the Pix2Pix-generated output (center), and the ground truth reference (right).

ID	SSIM ↑	PSNR (dB) ↑	MSE ↓
Depth	0.9543	38.45	9.36
NOCS	**0.9609**	**38.74**	**8.69**

TABLE I: Comparison of rendering metrics over the validation set of 1000 images for Depth and NOCS input modalities. Arrows indicate whether higher or lower values are better. Bold values highlight best performance of the NOCS input over all metrics.

A key limitation of path tracing rendering—particularly when combined with gradient computation through refractive media like glass—is its high computational cost. Although recent advances have enabled near real-time performance, path tracing remains significantly slower than data-driven or rasterization-based approaches, which often sidestep complex phenomena such as soft shadows, caustics, and refractions. To address this bottleneck, we propose using a traditional data-driven model like Pix2Pix, which achieves state-of-the-art inference speeds exceeding hundreds of frames per second and has demonstrated success in sim-to-real transfer for tactile sensors involving rigid bodies [12].

Our method also diverges from prior approaches in two critical ways. First, we eliminate the need for real data during training, instead using our refined simulation outputs as ground truth due to their photorealistic quality. Second, we move beyond depth maps, which inherently lack the ability to encode sensor translation and rotation. Depth alone cannot represent the full 6D pose information—including translational and rotational shear forces—that marker-based sensors can capture. To overcome this, we introduce NOCS (Normalized Object Coordinate Space [19]) maps as input. Commonly used in computer vision tasks such as shape reconstruction and category-level 6D pose estimation, NOCS maps encode per-pixel object geometry and pose. In our framework, these maps are derived from FEM mesh deformations, allowing them to implicitly capture translation and rotation within the mesh. This richer representation enhances learning and preserves high inference speed, despite the added complexity of modelling deformation and pose variation.

VI. Limitation and Future Works

While our approach shows promise, several avenues remain for further validation and extension. First, applying the method to other visual-tactile sensors could broaden its generalizability. Second, our framework is compatible with differentiable tactile simulation pipelines. Future work could explore integrating differentiable rendering techniques such as those proposed in DiffTactile [17], replacing marker-based differentiability with full RGB image gradients. This shift may be crucial for tasks requiring dense contact information, such as edge detection, shape reconstruction, or pose estimation, where sparse markers alone are insufficient.

VII. Conclusions

We have presented a comprehensive framework for high-fidelity simulation of visual tactile sensors. Our core innovation is a differentiable rendering pipeline that enables the optimization of physical parameters directly from real data, effectively closing the sim-to-real visual gap. By modeling complex lighting condition, we achieve unprecedented accuracy in simulating the appearance of markers under a protective glass layer matching the real tactile sensor. The optimized parameters from our pipeline produce a simulator that serves as a robust foundation for generating large, photorealistic datasets. To address the computational demands of path tracing, we introduced an efficient image-to-image translation model that uses rich NOCS maps as input, enabling real-time inference while preserving critical deformation information lost in depth-only representations. We validated our approach through benchmarking on a real robotic system, demonstrating superior performance over existing methods. We also showcased a novel application of our differentiable pipeline: solving the inverse problem of reconstructing a deformed mesh geometry from a single tactile image.

REFERENCES

[1] A. Agarwal, T. Man, and W. Yuan, "Simulation of vision-based tactile sensors using physics based rendering," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 1–7.

[2] I. Akinola, J. Xu, J. Carius, D. Fox, and Y. Narang, "Tacsl: A library for visuotactile sensor simulation and learning," *IEEE Transactions on Robotics*, 2025.

[3] W. Chen, J. Xu, F. Xiang, X. Yuan, H. Su, and R. Chen, "General-purpose sim2real protocol for learning contact-rich manipulation with marker-based visuotactile sensors," *IEEE Transactions on Robotics*, vol. 40, pp. 1509–1526, 2024.

[4] W. Chen, Y. Xu, Z. Chen, P. Zeng, R. Dang, R. Chen, and J. Xu, "Bidirectional sim-to-real transfer for gelsight tactile sensors with cyclegan," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6187–6194, 2022.

[5] Z. Chen, S. Zhang, S. Luo, F. Sun, and B. Fang, "Tacchi: A pluggable and low computational cost elastomer deformation simulator for optical tactile sensors," *IEEE Robotics and Automation Letters*, vol. 8, no. 3, pp. 1239–1246, 2023.

[6] S. Cui, Y. Wang, S. Wang, Q. Li, R. Wang, and C. Zhang, "Tactile imprint simulation of gelstereo visuotactile sensors," in *2023 IEEE International Conference on Mechatronics and Automation (ICMA)*, 2023, pp. 650–656.

[7] G. Duret, F. Zara, J. Peters, and L. Chen, "Toward synthetic data generation for robotic tactile manipulations," in *Workshop on "Robot Embodiment through Visuo-Tactile Perception" - 2024 IEEE International Conference on Robotics and Automation (ICRA) Conference Workshop*, Yokohama, Japan, May 2024. [Online]. Available: https://hal.science/hal-04566202

[8] D. F. Gomes, P. Paoletti, and S. Luo, "Generation of gelsight tactile images for sim2real learning," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 4177–4184, 2021.

[9] C. Higuera, B. Boots, and M. Mukadam, "Learning to read braille: Bridging the tactile reality gap with diffusion models," 04 2023.

[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017.

[11] W. Jakob, S. Speierer, N. Roussel, M. Nimier-David, D. Vicini, T. Zeltner, B. Nicolet, M. Crespo, V. Leroy, and Z. Zhang, "Mitsuba 3 renderer," 2022, https://mitsuba-renderer.org.

[12] Y. Lin, J. Lloyd, A. Church, and N. F. Lepora, "Tactile gym 2.0: Sim-to-real deep reinforcement learning for comparing low-cost high-resolution robot touch," ser. Proceedings of Machine Learning Research, R. L. A.Banerjee, Ed., vol. 7, no. 4. IEEE, August 2022, pp. 10 754–10 761. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9847020

[13] S. Luo, N. F. Lepora, W. Yuan, K. Althoefer, G. Cheng, and R. Dahiya, "Tactile robotics: An outlook," *arXiv preprint arXiv:2508.11261*, 2025.

[14] D. H. Nguyen, G. Duret, T. Schneider, A. Kshirsagar, B. Belousov, and J. Peters, "Tacex: Gelsight tactile simulation in isaac sim–combining soft-body and visuotactile simulators," in *CoRL Workshop on Learning Robot Fine and Dexterous Manipulation: Perception and Control*.

[15] T. Schneider, G. Duret, C. de Farias, R. Calandra, L. Chen, and J. Peters, "Tactile mnist: Benchmarking active tactile perception," *arXiv preprint arXiv:2506.06361*, 2025.

[16] Z. Si and W. Yuan, "Taxim: An example-based simulation model for gelsight tactile sensors," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2361–2368, 2022.

[17] Z. Si, G. Zhang, Q. Ben, B. Romero, Z. Xian, C. Liu, and C. Gan, "Difftactile: A physics-based differentiable tactile simulator for contact-rich robotic manipulation," in *The Twelfth International Conference on Learning Representations*.

[18] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, "Normalized object coordinate space for category-level 6d object pose and size estimation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[19] ——, "Normalized object coordinate space for category-level 6d object pose and size estimation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[20] S. Wang, M. Lambeta, P.-W. Chou, and R. Calandra, "Tacto: A fast, flexible, and open-source simulator for high-resolution vision-based tactile sensors," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3930–3937, 2022.

[21] J. Xu, S. Kim, T. Chen, A. R. Garcia, P. Agrawal, W. Matusik, and S. Sueda, "Efficient tactile simulation with differentiability for robotic manipulation," in *6th Annual Conference on Robot Learning*, 2022. [Online]. Available: https://openreview.net/forum?id=6BIffCl6gsM

[22] C. Zhang, S. Cui, J. Hu, T. Jiang, T. Zhang, R. Wang, and S. Wang, "Tacflex: Multi-mode tactile imprints simulation for visuotactile sensors with coating patterns," *IEEE Transactions on Robotics*, 2025.

[23] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.