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

TransTac: Visuo-Tactile Modality Transition via Ultraviolet-Encoded Transparent Elastomers

Lingyue Yang[1] and Bin Fang[1]

***Abstract*— Vision-based tactile sensors (VBTS) recover high-resolution contact geometry but typically rely on opaque elastomer layers that prevent visual transparency, while RGB-D cameras provide global depth perception yet degrade significantly at close range. To address this limitation, we present *TransTac*, a transparent ultraviolet (UV)-encoded binocular VBTS that integrates visual observation and marker-based tactile reconstruction within a single compact device. The system employs a transparent elastomer embedded with UV-reflective markers and a prior-guided Delaunay stereo matching algorithm for robust sparse triangulation.**

To reliably detect densely distributed semitransparent markers, we develop a lightweight detector that enables stable localization under contact and deformation. The proposed prior-guided Delaunay matching improves correspondence robustness by approximately 21% compared with global assignment baselines while maintaining high reconstruction accuracy. In semantic evaluation, TransTac achieves up to 83.3% zero-shot recognition accuracy on tactile images, exceeding opaque tactile baselines by approximately 50 percentage points. Embedding analysis further reveals substantially stronger cross-modal alignment with natural images, with class-center similarity increasing from around 0.2 to over 0.77. Controlled near-distance experiments quantify the degradation of RGB-D depth reliability and demonstrate extended geometric coverage enabled by visuo-tactile integration. Finally, a compact prototype is implemented with an approximate hardware cost of $70.

Code and hardware design are publicly available at https://github.com/87361/TransTac.

I. Introduction

Tactile sensing is essential for robotic manipulation, providing information about object shape [1], contact state [2], and material properties [3], [4] that vision alone cannot reliably capture. Vision-based perception often degrades under occlusion [5] or challenging illumination, preventing reliable estimation of contact geometry during manipulation. While tactile feedback can compensate for such failures, most existing robotic systems still lack compact and multimodal tactile sensors capable of delivering both accurate contact geometry and appearance-related cues at the interaction interface. This limitation restricts robust grasping and adaptive policy selection in unstructured environments.

This work was supported by the Brain Science and Brain-like Intelligence Technology - National Science and Technology Major Project (Grant No. 2025ZD0215600), the National Natural Science Foundation of China under Grant No.62573063, 62536001, the Open Foundation of the State Key Laboratory of Precision Space-time Information Sensing Technology No.STSL2025-B-07-01, and the Open Projects Program of State Key Laboratory of Multimodal Artificial Intelligence Systems.

[1]Beijing University of Posts and Telecommunications, Beijing, China. Email: yly-valentina@bupt.edu.cn, fangbin1120@bupt.edu.cn

Corresponding author: Bin Fang (fangbin1120@bupt.edu.cn).

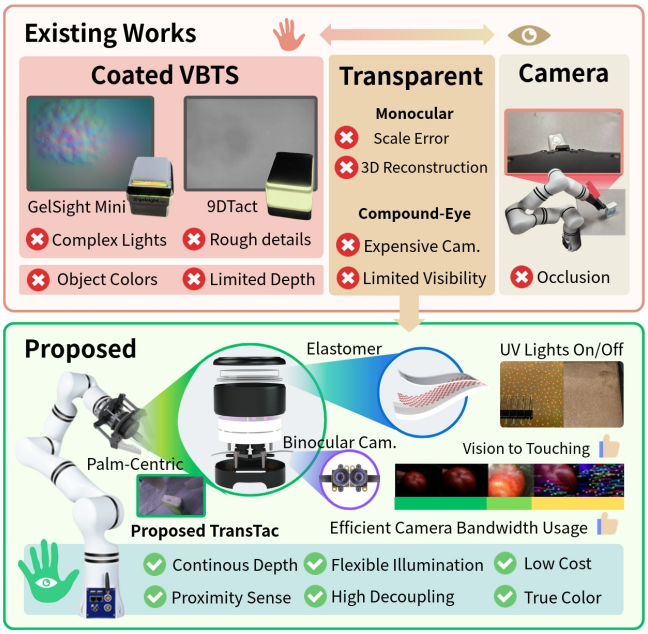


Fig. 1: Motivation of **TransTac**. RGB-D cameras provide global depth but degrade at close range, while coated VBTS reconstruct only contact deformation. TransTac integrates binocular vision and UV-encoded transparent tactile sensing within a unified hardware structure.

Existing vision-based tactile sensors (VBTS) are largely geometry-oriented. Systems such as GelSight [6] and its variants [7] reconstruct high-resolution contact deformation using coated elastomers, enabling tasks such as slip detection [8] and fine-grained shape reconstruction [9], [10]. However, the opaque reflective coating blocks visual observation through the sensing surface, requiring separate vision sensors for appearance and global geometry estimation.

The core **challenge** lies in maintaining reliable geometric perception across the transition from pre-contact observation to physical contact. RGB-D cameras provide global depth estimation but experience reduced depth reliability at close range, particularly near the minimum sensing distance [**?**]. In contrast, coated VBTS capture local deformation only after contact and cannot recover recessed or non-contact geometry. This separation between far-field depth estimation and near-field tactile reconstruction creates a geometric sensing gap.

As illustrated in Fig. 1, existing RGB-D perception and VBTS designs each address only part of this range. Bridging this gap requires a sensing configuration that preserves visual transparency, enables robust marker-based geometric

tracking, and supports metric depth recovery in near-contact scenarios.

In this work, we present **TransTac**, a transparent ultraviolet-encoded VBTS that integrates binocular RGB observation and marker-based tactile reconstruction within a compact stereo configuration. The transparent elastomer preserves visual clarity, while embedded UV-reflective markers enable robust geometric tracking under switchable illumination. We introduce a Delaunay-based stereo marker matching algorithm to establish reliable correspondences under epipolar constraints, enabling sparse metric triangulation of the contact surface. By combining real-time marker displacement tracking with stereo-based depth estimation and RGB-D fusion, TransTac supports both dynamic tactile interaction monitoring and near-contact geometric reconstruction.

Our contributions are summarized as follows:

- **A transparent ultraviolet-encoded binocular VBTS hardware design.** We present a stereo vision-based tactile sensor built upon a transparent elastomer embedded with UV-reflective markers, enabling simultaneous visual observation and marker-based tactile reconstruction within a single compact hardware structure.
- **A lightweight detection model for dense semitransparent markers.** We develop a compact marker detection network tailored for densely distributed semitransparent markers, achieving robust localization under varying illumination and contact conditions.
- **A prior-guided Delaunay stereo matching algorithm.** We introduce a Delaunay-based marker correspondence method initialized by epipolar nearest-neighbor matching, enabling reliable sparse triangulation and metric contact-surface reconstruction under geometric constraints.

II. Related Work

A. Vision-Based Tactile Sensing

VBTS recovers contact geometry by observing elastomer deformation with cameras. Representative systems such as GelSight [6] and the GelSlim series [2], [11] employ reflective coatings and photometric stereo to reconstruct high-resolution surface geometry. Learning-based mappings have further been introduced to infer contact shape from calibrated visual signals, as demonstrated by DIGIT [7] and the DTact family [12], [13]. While these designs achieve accurate local deformation sensing, they typically rely on opaque reflective layers and controlled illumination.

Transparent or vision-through tactile sensors have also been explored. Yamaguchi and Atkeson [14] demonstrated a transparent optical tactile skin capable of both deformation sensing and external visual observation. However, maintaining reliable geometric reconstruction while preserving visual transparency remains challenging, especially in near-contact scenarios where visual and tactile signals may interfere.

B. Marker-Based Tactile Reconstruction

Embedding markers within deformable membranes provides an alternative approach to tactile geometry estimation. Systems such as FingerVision [15], TacTip [16], and TAC2Pose [17] track marker displacement to infer contact information, while transparent implementations such as CompdVision [18] explore marker-based reconstruction under optical constraints.

Reliable marker detection under deformation and illumination variation remains challenging. Dense marker patterns may undergo geometric distortion during contact, affecting localization accuracy. Prior work has explored robust detection mechanisms [19] and illumination-based separation strategies such as UV-reflective markers in SpecTac [20]. Nevertheless, consistent detection and correspondence of dense semitransparent markers under switchable illumination remain difficult.

C. Stereo Correspondence and Near-Contact Depth Sensing

Reliable stereo correspondence is essential for triangulating marker positions in vision-based tactile sensing systems. Under rectified stereo geometry, correspondence search can be restricted to one-dimensional scanlines along epipolar lines. Classical approaches include global assignment methods such as Hungarian matching and support-point propagation strategies such as ELAS [21]. However, when applied to marker-based tactile sensing, correspondence estimation becomes more challenging due to repetitive marker appearance, dense spatial distribution, and deformation-induced geometric distortion.

In parallel, RGB-D sensing provides dense depth estimation but suffers from fundamental limitations in near-contact scenarios. Because of triangulation constraints and minimum sensing distance, the reliability of depth measurements degrades rapidly as objects approach the camera [22]. As a result, RGB-D sensors often fail to provide stable geometric cues in the immediate pre-contact region.

These limitations motivate combining tactile triangulation with visual depth estimation to achieve continuous geometric perception from pre-contact observation to physical contact.

III. Design and Fabrication

A. TransTac Design Principles

TransTac is designed to integrate visual observation and tactile reconstruction within a compact, low-cost stereo configuration. A transparent silicone membrane embedded with UV-reflective markers preserves optical clarity while enabling marker-based deformation sensing. Time-multiplexed illumination separates RGB imaging from UV marker observation without requiring additional optical components.

As illustrated in Fig. 2, binocular images are first rectified and processed for marker localization and stereo correspondence. Marker displacement provides local deformation cues, while triangulated correspondences enable sparse metric depth estimation. These components jointly support contact-surface reconstruction within a unified sensing interface.

B. Sensor Hardware and Fabrication

The sensing unit comprises a stereo pair of USB camera modules (HBVCAM-2436-R V11) with a horizontal field

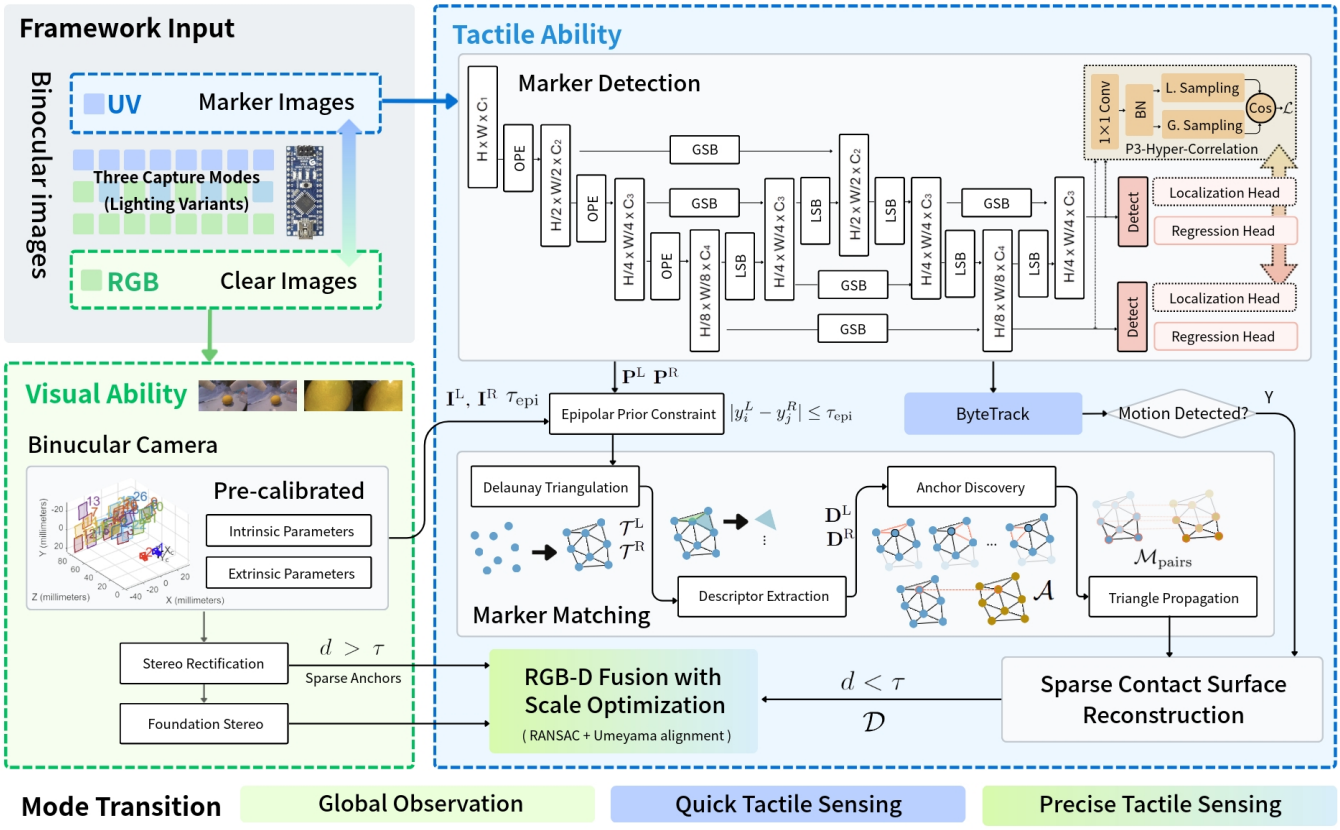


Fig. 2: Overview of the TransTac framework. Binocular RGB and UV images are rectified for marker detection and stereo correspondence. Sparse triangulated depth is fused with RGB-D estimation for contact-surface reconstruction. Color-coded modules indicate the final processing modules in each sensing stage.

of view of 100° and manual focus. Intrinsic and extrinsic parameters are obtained through standard multi-view calibration using a checkerboard of size GP052.

The tactile interface is fabricated from transparent silicone elastomer (Solaris™, 1:1 ratio) cast into a 3D-printed mold with a thickness of 5 mm. Fluorescent markers are manually embedded at irregular spatial locations and encapsulated within a thin silicone layer to reduce light attenuation. Illumination is provided by a 365 nm ultraviolet strip and a white LED strip.

The outer housing is 3D-printed in matte black material to suppress internal reflections and boundary artifacts.

C. Marker Detection and Tracking

Accurate localization of densely distributed semitransparent markers is critical for reliable stereo correspondence and deformation tracking. Compared with conventional blob-based methods, the proposed setting presents additional challenges, including marker translucency, UV/RGB illumination switching, and deformation-induced appearance variation.

a) Network architecture: As illustrated in Fig. 2, we adopt a lightweight single-stage, anchor-free detector built upon a compact convolutional backbone [23]. A correlation modeling module is introduced to capture both local and higher-order spatial relations among densely packed small markers. The detection head follows a CenterNet-style design [24], consisting of a Gaussian-encoded center heatmap branch and a regression branch for bounding box refinement. Skip connections are employed to preserve fine-grained spatial details.

b) Training objective: The overall loss is defined as

$$\mathcal{L} = w_{\text{loc}}\mathcal{L}_{\text{loc}} + w_{\text{reg}}\mathcal{L}_{\text{reg}} + w_{\text{corr}}\mathcal{L}_{\text{corr}}, \tag{1}$$

where the localization term supervises Gaussian-encoded center prediction, the regression term refines bounding box geometry, and the correlation term enforces feature consistency among markers within the same local cluster. This formulation improves robustness under deformation and partial occlusion without significantly increasing computational cost.

c) Tracking: Detected markers are associated across frames using ByteTrack [25] to maintain identity consistency and to estimate displacement for tactile deformation analysis.

D. Prior-Guided Stereo Marker Matching

Given rectified stereo images $\mathbf{I}^L$ and $\mathbf{I}^R$, let $\mathcal{P}^L = \{\mathbf{p}_i^L\}$ and $\mathcal{P}^R = \{\mathbf{p}_j^R\}$ denote the detected marker centroids in the left and right views. The objective is to establish reliable stereo correspondences $\mathcal{M} = \{(\mathbf{p}_i^L, \mathbf{p}_j^R)\}$ under epipolar and geometric constraints.

a) Epipolar Prior Initialization: For each marker $\mathbf{p}_i^L$, candidate matches in the right image are first restricted by the epipolar constraint

$$|y_i^L - y_j^R| \leq \tau_{\text{epi}}.$$

Among valid candidates, the nearest neighbor in disparity space is selected to form an initial correspondence set $\mathcal{M}_0$. These high-confidence matches provide statistical priors for subsequent anchor discovery.

b) Delaunay Encoding and Anchor Discovery: Delaunay triangulation is applied to $\mathcal{P}^L$ and $\mathcal{P}^R$, producing triangle sets $\mathcal{T}^L$ and $\mathcal{T}^R$. For a triangle t_{ijk} we define a normalized edge-length descriptor

$$\mathbf{d}_{ijk} = \frac{[\|\mathbf{p}_i - \mathbf{p}_j\|,\ \|\mathbf{p}_j - \mathbf{p}_k\|,\ \|\mathbf{p}_k - \mathbf{p}_i\|]}{\|[\|\mathbf{p}_i - \mathbf{p}_j\|, \|\mathbf{p}_j - \mathbf{p}_k\|, \|\mathbf{p}_k - \mathbf{p}_i\|]\|_2}.$$

Triangle pairs satisfying descriptor similarity and epipolar consistency are selected as anchor seeds $\mathcal{A}$.

c) Triangle Propagation: Starting from anchor pairs, correspondences are propagated to adjacent triangles while enforcing geometric consistency. This local structural propagation preserves neighborhood topology under moderate deformation and avoids ambiguity in global assignment.

The resulting set $\mathcal{M}$ provides robust stereo correspondences for subsequent sparse triangulation and dense depth alignment.

E. RGB-D Fusion with Scale Optimization

The sparse triangulated marker points provide reliable geometric depth near the contact region, while dense depth is estimated from binocular RGB images using the pretrained stereo model FoundationStereo [26]. However, the predicted dense depth may suffer from scale bias. To combine the advantages of both sources, we align the dense depth map with geometrically reliable sparse triangulated depths.

a) Distance-Aware Sparse Anchors: The switching threshold τ is defined as the calibrated distance between the stereo cameras and the elastomer surface. For $d > \tau$, visual stereo features provide sparse anchors, while for $d < \tau$ triangulated markers with detectable displacement (tracked by ByteTrack) are used instead.

b) Sparse Triangulation: Given matched marker pairs (p_i^L, p_i^R) and calibrated stereo projection matrices

$$P_L = K_L[I|0], \quad P_R = K_R[R|t],$$

each correspondence is triangulated to obtain a sparse 3D point

$$X_i^{\text{sparse}} = (X_i, Y_i, Z_i).$$

These triangulated points provide metrically reliable depth observations near the contact region.

c) Dense Depth Alignment: Meanwhile the RGB-D model predicts a dense depth map $D_{net}(u, v)$. Each pixel is back-projected into a dense 3D point

$$X_{dense}(u, v) = \left(\frac{u - c_x}{f_x} Z, \frac{v - c_y}{f_y} Z, Z\right), \quad Z = D_{net}(u, v).$$

For each triangulated feature location (u_i, v_i) we obtain the corresponding dense point X_i^{dense}, producing sparse–dense correspondences $(X_i^{sparse}, X_i^{dense})$.

After removing outliers using RANSAC, a similarity transformation

$$T = (s, R, t)$$

is estimated via Umeyama alignment to minimize

$$\sum_i \|X_i^{sparse} - (sRX_i^{dense} + t)\|^2.$$

The transformation is then applied to all dense points to obtain a metrically consistent dense depth map.

IV. Experiments and Results

We evaluate the proposed TransTac system across four key capabilities: (i) semantic recognizability of tactile imagery, (ii) recovery of near-contact geometry in regions where RGB-D sensing fails, (iii) robustness of stereo marker correspondence estimation, and (iv) stability of tactile marker tracking during interaction.

Experiments are designed to analyze both perceptual and geometric aspects of the proposed visuo–tactile sensing framework. Semantic evaluation focuses on whether tactile observations preserve high-level visual semantics that can be interpreted by modern vision-language models. Geometric evaluation investigates the system's ability to recover accurate contact geometry in near-contact scenarios where conventional RGB-D sensors exhibit depth degradation.

For experiments involving geometric reconstruction, ground-truth geometry is obtained from STL models of 3D-printed objects. This provides sub-millimeter reference accuracy without relying on noisy active depth measurements.

For quantitative experiments, the sensor is rigidly mounted on a three-axis gantry system that provides repeatable positioning with sub-millimeter precision along the Z-axis. This setup enables controlled evaluation of depth estimation and reconstruction stability at varying distances. For qualitative experiments, the sensor is handheld to allow diverse contact poses, sliding interactions, and realistic manipulation scenarios.

A. Semantic Recognizability of Tactile Images

We first evaluate whether tactile images captured by TransTac preserve sufficient visual semantics for object-level recognition. Two vision-language models (Qwen-VLM and ChatGPT-VLM) and two open-vocabulary detectors (YOLO-World [27] and YOLO-E [28]) are used to assess semantic interpretability across tactile modalities.

TABLE I: Semantic preservability across tactile modalities. All recognition models are evaluated using a weighted semantic scoring scheme.

Model	GelSight	9DTact	TransTac
Recognition			
Qwen-VLM	28.7	10.8	**80.6**
ChatGPT-VLM	30.2	12.5	**83.3**
YOLO-World	15.3	2.1	**72.2**
YOLO-E	16.9	3.6	**75.0**
Embedding			
SigLIP2 (Zero-shot %)	27.8	27.8	**83.3**
DINOv2 (Center Similarity)	0.236	0.202	**0.774**
DINOv2 (NN Top-1 %)	27.8	55.6	**100.0**

Three sensing modalities are considered: GelSight Mini, 9DTact, and the proposed TransTac sensor. Six object categories (egg, coin, battery, Lego block, button, and glass bead) are used, with two instances per category. Each object is captured from three viewpoints, producing 36 tactile images for each modality.

Since tactile images may contain partial object cues, recognition performance is evaluated using a *weighted semantic scoring scheme*. Predictions that partially match the ground-truth category receive proportional credit, allowing fair comparison across sensing modalities and models.

Table I summarizes both recognition results and embedding-level alignment. Across all recognition models, TransTac consistently achieves substantially higher semantic scores than opaque tactile sensors. For example, ChatGPT-VLM achieves 83.3% on TransTac images, compared with 30.2% for GelSight and 12.5% for 9DTact.

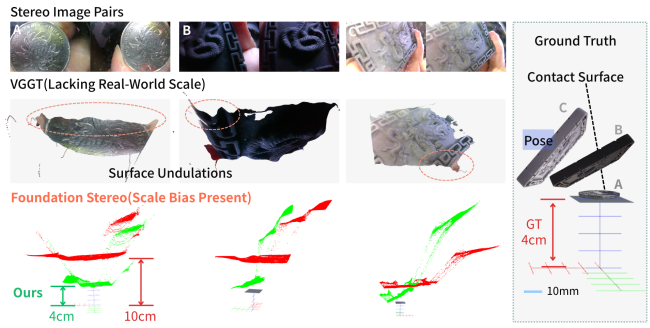
To further analyze representation-level consistency, we compute embedding similarity using SigLIP2 [29] and DINOv2 [30]. TransTac tactile images remain significantly closer to natural image representations, with DINOv2 class-center similarity increasing from approximately 0.20–0.24 for opaque tactile sensors to 0.774 for TransTac. Nearest-neighbor alignment also reaches 100%, indicating strong cross-modal consistency between tactile and visual representations.

B. Near-Contact Geometry Recovery

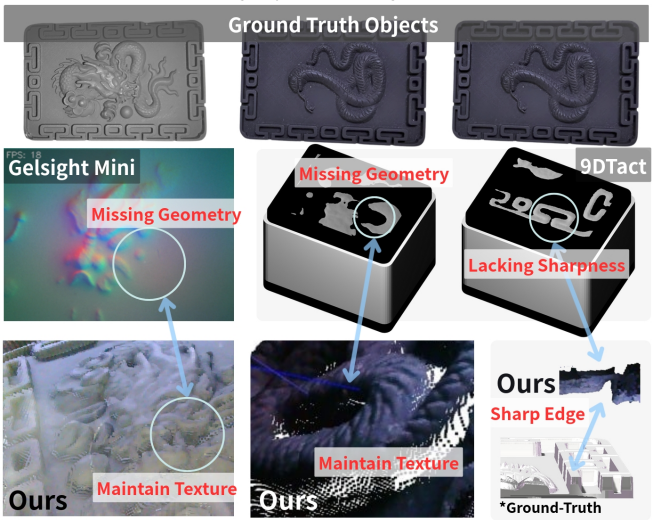
Most RGB-D sensors (e.g., Intel RealSense and Microsoft Kinect [**?**], [31]) have minimum sensing distances that extend beyond the near-contact region. Even sensors capable of closer-range operation exhibit rapid degradation as objects approach the camera, leaving a sensing gap immediately before physical contact.

Coated vision-based tactile sensors (VBTS) suffer from a different limitation. Because the sensing layer is opaque and planar, they can only observe the contact interface and cannot capture recessed surface geometry [9], [32].

To illustrate these limitations, we compare TransTac with both RGB-D depth estimation methods and coated VBTS sensors. Fig. 3(a) shows that recent depth estimation models such as VGGT [33] and FoundationStereo [26] struggle in



(a) Scale ambiguity in learning-based methods.



(b) Limitations of coated VBTS in perceiving recessed geometry.

Fig. 3: Qualitative comparison of contact surface reconstruction.

this setting. VGGT fails to recover the correct metric scale, interpreting perspective cues as planar projections, while FoundationStereo preserves relative structure but exhibits absolute scale errors due to the short stereo baseline and distribution mismatch.

Fig. 3(b) compares our approach with coated VBTS sensors. These sensors exhibit two typical artifacts: (i) recessed geometries are missing because only protruding regions contact the sensing surface, and (ii) elastic deformation smooths sharp edges into slopes, producing false depths along object boundaries.

To quantify the near-contact limitation of RGB-D sensing, we measure the valid depth ratio of an Intel RealSense D405 as the distance between the sensor and a planar object decreases. The valid depth ratio is defined as the proportion of pixels with non-zero depth values within the region of interest.

Fig. 4(a) plots the valid depth ratio as a function of physical distance. Depth coverage remains relatively stable when the object is farther than approximately 9 cm from the sensor. As the object enters the near-contact region, however, the proportion of valid depth pixels drops sharply and eventually

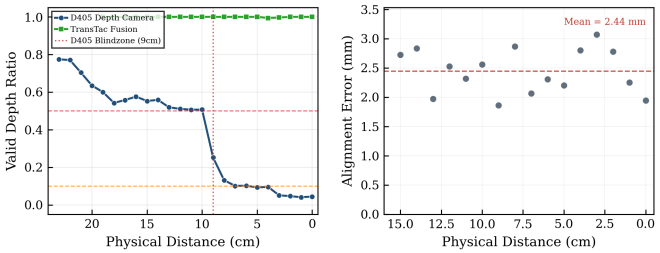


Fig. 4: Near-contact depth sensing and alignment evaluation. (a) Valid depth ratio of an Intel RealSense D405 versus distance. (b) Sparse–dense alignment error across distances after RANSAC filtering.

falls below 10%, revealing a near-contact sensing blind zone where RGB-D sensing becomes unreliable.

Despite this degradation, sparse triangulated markers still provide stable geometric constraints. Fig. 4(b) shows the alignment error between triangulated marker depths and dense depth predictions after scale alignment. Across all evaluated distances, the mean geometric error remains approximately $2.44\,\mathrm{mm}$, indicating that reliable geometric cues can still be obtained in the near-contact region.

Overall, TransTac combines RGB-D estimation with marker-based triangulation to provide continuous geometric observations from approach to physical contact.

C. Stereo Marker Matching Robustness

Reliable stereo marker correspondence is critical for accurate triangulation of marker positions. However, dense semi-transparent markers embedded in a deformable elastomer introduce several challenges for correspondence estimation, including repetitive appearance, dense spatial distribution, and geometric distortion caused by elastomer deformation during contact.

To evaluate the robustness of the proposed matching strategy, we compare the prior-guided Delaunay matching with several baseline approaches. The evaluated baselines include epipolar nearest-neighbor matching and global assignment using the Hungarian algorithm. We also include representative dense correspondence methods, including optical flow (Farneback) and stereo block matching (SGBM), for comparison.

The benchmark is constructed using stereo marker observations collected under diverse interaction conditions, including varying contact pressures, sliding motions, and partial occlusions. Ground-truth correspondences are manually annotated for evaluation.

Since the stereo cameras observe the transparent elastomer layer from slightly different viewpoints, a small portion of markers near the image boundaries may appear in only one view. In addition, the markers are randomly distributed across the sensing surface, so the set of markers visible in the two images is not perfectly identical. As a result, the number of markers that can be matched between the stereo pair varies slightly across frames.

TABLE II: Stereo marker correspondence comparison under different matching strategies. Results report the average number of correctly matched markers.

Method	Avg. Correct Matches
Marker-Based Matching	
Prior-Guided Delaunay (Ours)	**90.8**
Hungarian Assignment	74.9
Epipolar Nearest Neighbor	74.5
Dense Correspondence	
Optical Flow (Farneback)	37.9
Stereo Matching (SGBM)	28.7

Table II reports the average number of correctly matched markers. The proposed prior-guided Delaunay matching achieves the best performance with 90.8 correct matches on average, substantially improving over Hungarian assignment with 74.9 matches and epipolar nearest neighbor with 74.5 matches.

Dense correspondence methods perform noticeably worse in this scenario. Optical flow and stereo matching obtain 37.9 and 28.7 correct matches respectively. This behavior is mainly caused by the near-contact imaging geometry, where the left and right cameras observe different object surfaces. Regions visible in one view may correspond to different physical surfaces in the other view, reducing the number of reliable dense correspondences.

D. Tactile Marker Tracking

To assess tactile sensing quality, we track markers embedded beneath the transparent silicone layer of the sensor under diverse interaction conditions including varying contact pressure, tangential slip, rolling motion, and large deformation. The proposed method is compared with a dense optical flow baseline that has been widely used in prior work. The proposed detection model operates at 20 FPS in real time on a laptop GPU with NVIDIA RTX 4060 architecture.

We benchmark our marker detection capability against the traditional approach of color space thresholding combined with blob detection, which was employed in prior work utilizing UV-activated markers [20]. This conventional method exhibits several limitations, including frequent missed detections under varying illumination conditions or partial occlusion, as well as inconsistent marker identification and localization across frames.

The optical flow baseline computes pixel displacements between consecutive frames, making it susceptible to cumulative drift and unable to reliably separate tactile-induced motion from background visual changes. In contrast, our method applies displacement estimation, isolating marker movement associated with actual contact deformation. This decoupling prevents false positives from background motion or viewpoint changes. In non-contact cases, occasional specular highlights are filtered out, as they do not correspond to physically adhered contact points.

Qualitative sequences in Fig. 5 illustrate the difference:

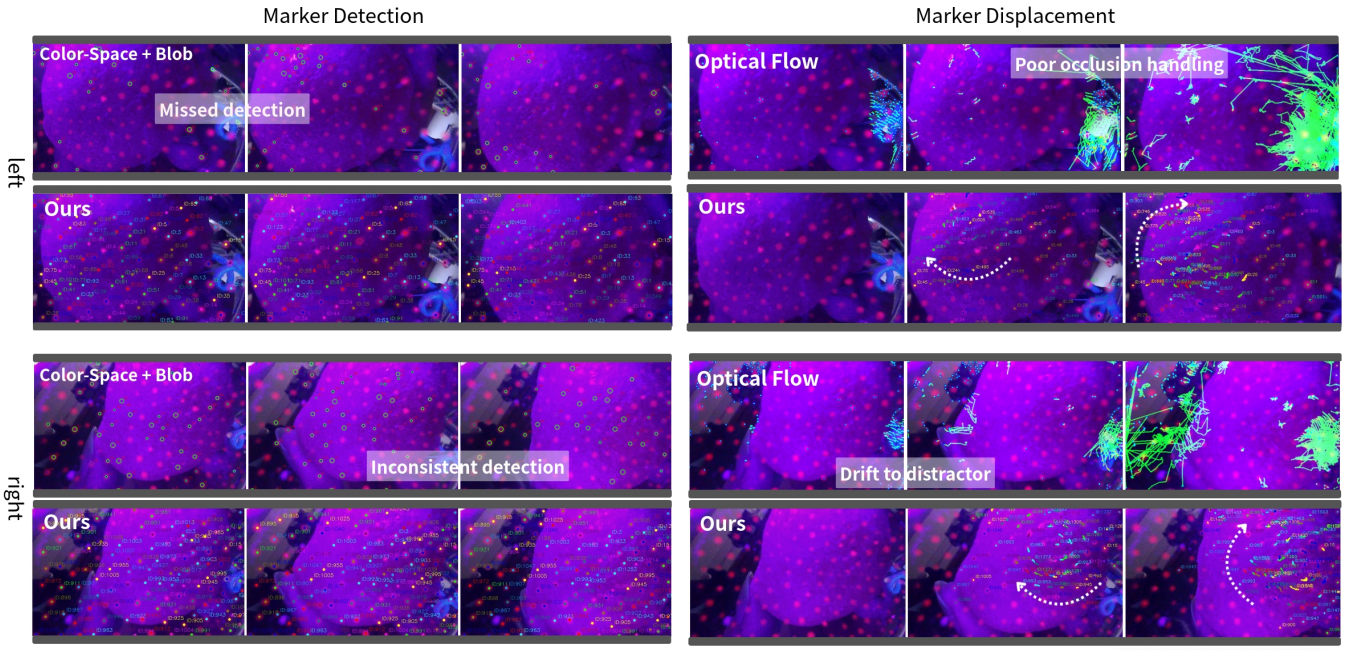


Fig. 5: Qualitative comparison: Left) Marker detection: traditional methods (prone to missed/inconsistent detections) vs our robust model. Right) Trajectory stability under slip/deformation: optical flow (drift/identity swaps) vs our stable tracking.

our system maintains stable marker identities and trajectories throughout challenging slip and deformation scenarios, whereas the optical flow baseline exhibits rapid drift, identity swaps, and intermittent loss of correspondences.

V. Conclusion

We presented TransTac, a transparent ultraviolet-encoded binocular vision-based tactile sensor that integrates marker-based tactile reconstruction and stereo observation within a compact hardware structure. The system combines semi-transparent marker detection, prior-guided Delaunay stereo matching, and sparse triangulation with RGB-D fusion to support contact-surface reconstruction and dynamic marker tracking.

Extensive experiments validate the effectiveness of the proposed system. In semantic evaluation, TransTac achieves up to 83.3% zero-shot recognition accuracy on tactile images across the evaluated models. Embedding-level analysis further confirms semantic preservation. The DINOv2 center similarity increases from around 0.2 for opaque VBTS to 0.774 with TransTac, indicating substantially improved alignment with natural image representations.

In stereo matching benchmarks, the proposed prior-guided Delaunay method improves correspondence robustness by approximately 21% compared with global assignment baselines while maintaining high matching accuracy. Controlled near-distance measurements quantify the degradation of RGB-D depth validity at close range and further show that the proposed visuo-tactile integration extends reliable geometric sensing coverage in near-contact scenarios.

Rather than replacing conventional vision or tactile systems, TransTac provides a complementary sensing configuration that preserves visual transparency while enabling metric contact reconstruction and appearance consistency within a unified device.

VI. Limitation and Future Work

Although TransTac enables integrated visual observation and tactile reconstruction, several limitations remain. First, the current system focuses primarily on geometric sensing and does not yet directly measure force or pressure distributions. Extending the sensor to support force estimation through elastomer deformation modeling or embedded sensing elements is an important direction for future work.

Second, the detection and tracking of densely distributed markers require annotated training data, and collecting labeled datasets for marker-based tactile sensing remains labor-intensive. In addition, existing depth estimation models are typically trained on datasets with limited extreme close-range or manipulation scenarios, which may lead to scale inaccuracies in near-contact depth recovery. Future work will explore improved data collection strategies and domain adaptation techniques for near-contact perception.

Third, the current prototype uses off-the-shelf stereo camera modules and illumination components. Hardware optimization, including smaller camera modules and integrated fingertip-scale designs, could further improve compactness and enable deployment on robotic grippers or dexterous hands. In addition, switching between illumination modes introduces minor latency due to camera exposure adaptation, which could be reduced through tighter hardware–software integration.

Finally, the fabrication process of ultraviolet fluorescent markers currently relies on manual or semi-automatic procedures, and the durability of the fluorescent materials under long-term use remains to be systematically evaluated.

Future work will investigate automated marker deposition techniques and more robust fluorescent materials to improve manufacturability and long-term stability.

ACKNOWLEDGMENTS

The authors acknowledge the use of generative AI tools for language polishing and minor code prototyping assistance. All experimental design, implementation, analysis, and conclusions are solely the responsibility of the authors.

REFERENCES

[1] Q. Li and S. Yuan, "Jacquard v2: Refining datasets using the human in the loop data correction method," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 7932–7938.

[2] E. Donlon, S. Dong, M. Liu, J. Li, E. Adelson, and A. Rodriguez, "Gelslim: A high-resolution, compact, robust, and calibrated tactile-sensing finger," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1927–1934.

[3] J. Zheng, J. Zhang, K. Yang, K. Peng, and R. Stiefelhagen, "Materobot: Material recognition in wearable robotics for people with visual impairments," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 2303–2309.

[4] S. Zhang, Y. Sun, J. Shan, Z. Chen, F. Sun, Y. Yang, and B. Fang, "Tirgel: A visuo-tactile sensor with total internal reflection mechanism for external observation and contact detection," *IEEE Robotics and Automation Letters*, vol. 8, no. 10, pp. 6307–6314, 2023.

[5] M. A. Lee, Y. Zhu, P. Zachares, M. Tan, K. Srinivasan, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Learning multimodal representations for contact-rich tasks," *IEEE Transactions on Robotics*, vol. 36, no. 3, pp. 582–596, 2020.

[6] 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.

[7] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. 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*, vol. 5, no. 3, pp. 3838–3845, 2020.

[8] J. W. James, N. Pestell, and N. F. Lepora, "Slip detection with a biomimetic tactile sensor," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3340–3346, 2018.

[9] C. Lu, K. Tang, X. Hui, H. Li, S. Nam, and N. F. Lepora, "Stereotactip: Vision-based tactile sensing with biomimetic skin-marker arrangements," *IEEE/ASME Transactions on Mechatronics*, 2025.

[10] W. K. Do and M. Kennedy, "Densetact: Optical tactile sensor for dense shape reconstruction," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 6188–6194.

[11] I. H. Taylor, S. Dong, and A. Rodriguez, "Gelslim 3.0: High-resolution measurement of shape, force and slip in a compact tactile-sensing finger," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 10 781–10 787.

[12] C. Lin, Z. Lin, S. Wang, and H. Xu, "Dtact: A vision-based tactile sensor that measures high-resolution 3d geometry directly from darkness," *arXiv preprint arXiv:2209.13916*, 2022.

[13] C. Lin, H. Zhang, J. Xu, L. Wu, and H. Xu, "9dtact: A compact vision-based tactile sensor for accurate 3d shape reconstruction and generalizable 6d force estimation," *IEEE Robotics and Automation Letters*, vol. 9, no. 2, pp. 923–930, 2023.

[14] A. Yamaguchi and C. G. Atkeson, "Combining finger vision and optical tactile sensing: Reducing and handling errors while cutting vegetables," in *2016 IEEE-RAS 16th international conference on humanoid robots (humanoids)*. IEEE, 2016, pp. 1045–1051.

[15] Y. Zhang, Z. Kan, Y. A. Tse, Y. Yang, and M. Y. Wang, "Fingervision tactile sensor design and slip detection using convolutional lstm network," *arXiv preprint arXiv:1810.02653*, 2018.

[16] B. Ward-Cherrier, N. Pestell, L. Cramphorn, B. Winstone, M. E. Giannaccini, J. Rossiter, and N. F. Lepora, "The tactip family: Soft optical tactile sensors with 3d-printed biomimetic morphologies," *Soft robotics*, vol. 5, no. 2, pp. 216–227, 2018.

[17] M. Bauza, A. Bronars, and A. Rodriguez, "Tac2pose: Tactile object pose estimation from the first touch," *The International Journal of Robotics Research*, vol. 42, no. 13, pp. 1185–1209, 2023.

[18] L. Luo, B. Zhang, Z. Peng, Y. K. Cheung, G. Zhang, Z. Li, M. Y. Wang, and H. Yu, "Compdvision: Combining near-field 3d visual and tactile sensing using a compact compound-eye imaging system," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 262–268.

[19] M. Li, Y. H. Zhou, T. Li, and Y. Jiang, "Real-time and robust feature detection of continuous marker pattern for dense 3-d deformation measurement," *Measurement*, vol. 221, p. 113479, 2023.

[20] Q. Wang, Y. Du, and M. Y. Wang, "Spectac: A visual-tactile dual-modality sensor using uv illumination," in *2022 international conference on robotics and automation (ICRA)*. IEEE, 2022, pp. 10 844–10 850.

[21] A. Geiger, M. Roser, and R. Urtasun, "Efficient large-scale stereo matching," in *Asian conference on computer vision*. Springer, 2010, pp. 25–38.

[22] M. Servi, E. Mussi, A. Profili, R. Furferi, Y. Volpe, L. Governi, and F. Buonamici, "Metrological characterization and comparison of d415, d455, l515 realsense devices in the close range," *Sensors*, vol. 21, no. 22, p. 7770, 2021.

[23] Y. Zhou, L. Li, L. Lu, and M. Xu, "nnwnet: Rethinking the use of transformers in biomedical image segmentation and calling for a unified evaluation benchmark," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 20 852–20 862.

[24] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6569–6578.

[25] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "Bytetrack: Multi-object tracking by associating every detection box," in *European conference on computer vision*. Springer, 2022, pp. 1–21.

[26] B. Wen, M. Trepte, J. Aribido, J. Kautz, O. Gallo, and S. Birchfield, "Foundationstereo: Zero-shot stereo matching," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5249–5260.

[27] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "Yolo-world: Real-time open-vocabulary object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 16 901–16 911.

[28] A. Wang, L. Liu, H. Chen, Z. Lin, J. Han, and G. Ding, "Yoloe: Real-time seeing anything," *arXiv preprint arXiv:2503.07465*, 2025.

[29] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa *et al.*, "Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," *arXiv preprint arXiv:2502.14786*, 2025.

[30] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[31] J. Smisek, M. Jancosek, and T. Pajdla, "3d with kinect," in *2011 IEEE international conference on computer vision workshops (ICCV Workshops)*. IEEE, 2011, pp. 1154–1160.

[32] Y. Sun, N. Cheng, S. Zhang, W. Li, L. Yang, S. Cui, H. Liu, F. Sun, J. Zhang, D. Guo *et al.*, "Tactile data generation and applications based on visuo-tactile sensors: A review," *Information Fusion*, vol. 121, p. 103162, 2025.

[33] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "Vggt: Visual geometry grounded transformer," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5294–5306.