DeTac: Dual-branch Decoupling of Texture and Force Signals in Vision-Based Tactile Sensors

Zhirui Zhang[1,2] **Yunkai Tang**[1,2] **Jiahang Li**[1,2] **Fan Fei**[1,2] **Jinnuo Zhang**[3]
Kairan Ding[1,2] **Jieji Ren**[3,†] **Yandong Guo**[2] **Boxin Shi**[1,†]
[1]Peking University [2]AI[2] Robotics [3]Shanghai Jiao Tong University
zrzhang25@stu.pku.edu.cn shiboxin@pku.edu.cn

Abstract: Vision-based tactile sensors enable fine-grained perception for dexterous robotic manipulation, but current marker-based designs tightly couple surface texture and contact force signals within a single observation. Existing data-driven decoupling methods rely on serial mask-based inpainting, which suffers from mask-sizing dilemmas caused by physical indentation effects, error retention, and lack of real paired data for supervision. To address these challenges, we propose DeTac, a novel dual-branch generative framework that concurrently extracts clean shading and marker images from a single mixed tactile image. DeTac leverages full-image synthesis to naturally bypass mask dependency and introduces a cross-branch synergistic optimization that enables mutual supervision, effectively preventing error retention. Furthermore, we construct a strictly aligned, real-world paired tactile dataset using custom hardware to provide high-quality supervision. Extensive experiments demonstrate that DeTac achieves state-of-the-art decoupling accuracy, significantly facilitating downstream surface reconstruction and force estimation tasks while exhibiting cross-sensor generalization.

Keywords: Vision-based tactile sensor, robot tactile sensing, image generation

1 Introduction

Vision-based tactile sensors [1] transform local contact signals into visible elastomer deformations under structured light, providing robots with fine-grained tactile perception crucial for dexterous manipulation [2, 3, 4, 5, 6, 7, 8]. While markerless sensors capture high-resolution shading images to reconstruct surface geometric textures [9, 10], their lack of spatial references hinders direct contact force estimation [11]. To address this, marker-based sensors embed physical markers into the elastomer [11, 12], tracking their displacements to calculate force [13, 14]. However, this design tightly *couples* texture and force signals within a single observation, as the markers inherently obstruct the surface shading that encodes geometric textures [15, 16]. Consequently, accurately *decoupling* these signals to reliably recover both surface texture and contact force has become a critical challenge.

While early attempts at signal decoupling relied on complex hardware for mode switching [17], the rise of generative models [18, 19, 20] has enabled data-driven decoupling directly from a single mixed image. Existing data-driven methods [21, 22] typically employ a *serial* pipeline: they first estimate a mask to locate the markers, then apply image inpainting [23, 24] to infer from surrounding regions the occluded shading region as if the markers did not exist. However, we observe that the physical embedding of markers, whether via sprayed ink [9] or 3D-printed pins [25], not only occludes the regions beneath the markers but also induces *indentation artifacts* (Fig. 1(a) vs. (b)): It alters the surface normals of the surrounding elastomer and therefore disrupts the perceived geometry texture at these regions. This phenomenon results in three major limitations in current pipelines (Fig. 1(c)). First, mask-based inpainting is confused by **mask-sizing dilemma**: Failing to

[†]Corresponding authors. [1]Zhirui Zhang, Yunkai Tang, Jiahang Li, Fan Fei, Kairan Ding, and Boxin Shi are with State Key Laboratory of Multimedia Information Processing, National Engineering Research Center of Visual Technology, and PKU-AI[2] Robotics Joint Lab of Embodied AI, School of Computer Science, Peking University. [3]Jinnuo Zhang and Jieji Ren are with the State Key Laboratory of Mechanical System and Vibration, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, and also with Shanghai Key Laboratory of Intelligent Robotics, Shanghai Jiao Tong University, Shanghai.

10th Conference on Robot Learning (CoRL 2026), Austin, Texas, USA.

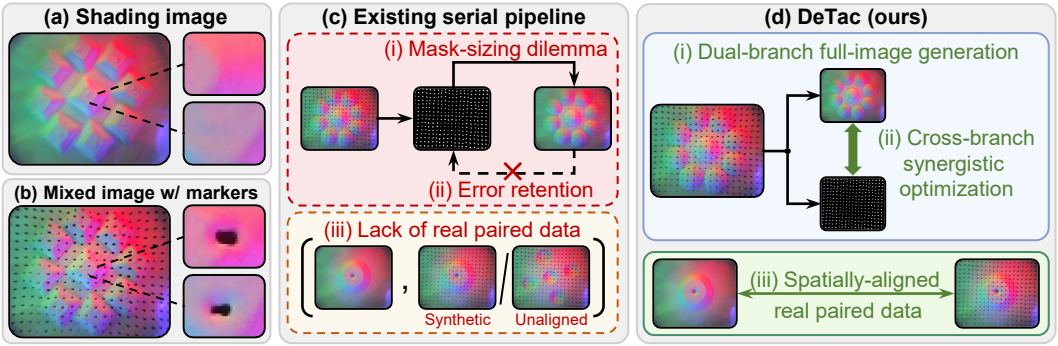


Figure 1: (a) Markerless sensors capture shading images to reconstruct surface but lack spatial references for force estimation. (b) Marker-based sensors enable force sensing via embedded markers, but introduce coupled signals and indentation artifacts. (c) Serial pipelines suffer from mask-size dilemmas, error retention, and lack of real paired training data. (d) DeTac uses a dual-branch framework with cross-branch optimization, trained on a self-collected aligned paired real dataset.

cover the altered-normal regions (which vary with force and position) corrupts the predicted shading, while oversized masks discard critical texture information [26, 27]. Second, serial prediction suffers from **error retention**: Initial errors in marker localization propagate to the shading restoration step without any feedback mechanism to correct them. Third, training is bottlenecked by the **lack of real paired data**: As strict alignment is difficult given real sensor fabrication variances, prior works rely on either synthetic marker superimposition that ignores indentation artifacts [21] or cycle-consistency using unaligned data [28, 29], both yielding suboptimal real-world performance.

To overcome these limitations, we propose DeTac, a *dual-branch* framework that Decouples mixed Tactile observations to a marker-free shading image and a marker image (Fig. 1(d)). Unlike mask-based inpainting, DeTac treats shading restoration and marker extraction as **concurrent full-image generation** tasks, adaptively modifying regions without relying on predefined masks, thus naturally bypassing the mask-sizing dilemma. We further introduce a **cross-branch synergistic optimization** that connects the two branches by enforcing spatial consistency between the predicted marker image and the restoration-induced shading changes, enabling mutual supervision that refines both predictions. Finally, to provide high-quality supervision, we construct a **real-world and strictly aligned paired dataset** using a specialized mechanical structure that captures exactly matched data on the same sensor before and after marker fabrication. In summary, our main contributions are:

- We propose DeTac, a novel dual-branch generative framework that uses full-image synthesis to decouple mixed tactile signals, jointly predicting clean shading and marker images.
- We introduce a cross-branch synergistic optimization that enables mutual supervision between shading restoration and marker extraction, effectively preventing error retention.
- We construct a strictly aligned, real-world paired tactile dataset using a specialized mechanical structure, providing high-quality supervision for data-driven methods.

Extensive experiments demonstrate that DeTac achieves state-of-the-art accuracy in both shading restoration (with an over 7 dB improvement in PSNR over TacDiff [21]) and marker extraction. Furthermore, the precisely decoupled signals significantly facilitate downstream tasks, including surface reconstruction and force estimation, while exhibiting generalization capabilities across tactile sensors.

2 Related Work

2.1 Vision-based Tactile Sensors

Vision-based tactile sensors [9, 10, 30, 25] provide detailed tactile feedback by transforming local contact information into visible deformations of a soft, transparent elastomer [31]. Early markerless designs [1, 32, 33] employ structured light to capture high-frequency geometric details of the contact interface, thereby preserving rich textural information. Marker-based sensors introduce dedicated fiducial markers to complement these designs with force estimation capabilities. GelSight [9, 34,

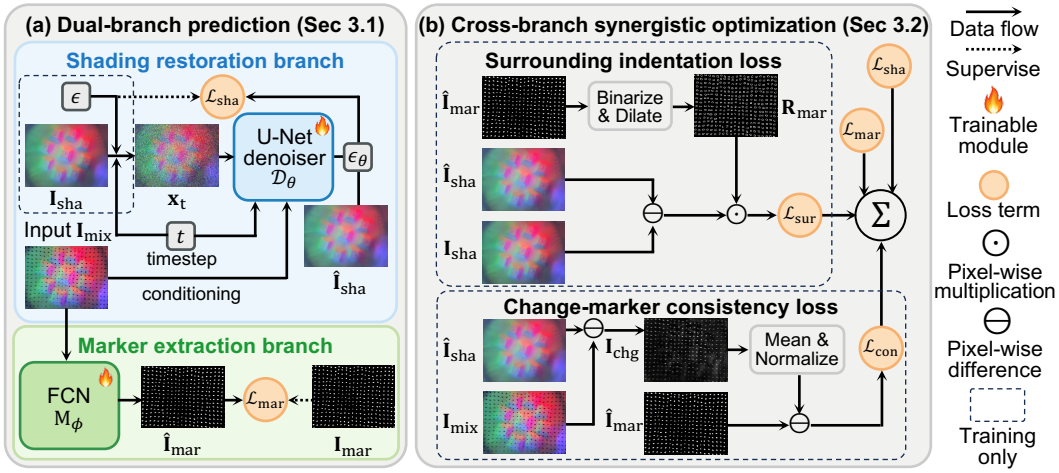


Figure 2: Overall pipeline of DeTac. (a) DeTac takes as input a mixed tactile image and consists of two parallel branches (Sec. 3.1): a conditional diffusion-based shading restoration branch and a lightweight marker extraction branch. (b) The cross-branch synergistic optimization (Sec. 3.2) exploits marker-aware spatial structures for mutual supervision using surrounding indentation loss for shading restoration and change-marker consistency loss for marker extraction.

11], GelSlim [12, 14], and DIGIT [30] fabricate markers by spraying or transferring 2D black ink arrays onto the elastomer surface, while the TacTip family [35, 36, 37, 25] utilizes dual-material 3D printing technology to monolithically manufacture biomimetic internal pins. Due to the indentation artifacts introduced in Sec. 1, the resulting sensor output is a complex mixture of shading and marker, where texture and force signals are inherently conflated. UVTac [17] addresses this by physically separating the two signals at the hardware output level, but introduces complex hardware overhead and limits the sensor to capturing only one modality at a time. TacDiff [21] attempts to decouple the two signals computationally, but is constrained by serial pipeline architectures and the scarcity of paired ground-truth (GT) data. These limitations motivate our dual-branch framework, which jointly recovers both modalities via cross-branch supervision, trained on a highly aligned paired dataset.

2.2 Conditional Image-to-image Generation

Conditional image-to-image generation synthesizes a target image guided by dense spatial or structural cues from a source image. Early work tackled this through generative adversarial network (GAN) [38]-based frameworks [24, 39], which framed the task as a direct domain mapping problem. However, due to the training instabilities inherent to GANs [38], recent approaches have largely shifted toward diffusion models [18]. Among these, masked inpainting methods regenerate only specific regions by conditioning on the unmasked surrounding context, either by concatenating the masked image and mask as additional input channels [40] or by employing a separate feature extractor [41]. However, these methods require the undesired and desired signal components to be strictly separable at pixel level; otherwise, information loss or quality degradation is inevitable. Full-image generation methods avoid this limitation by generating the complete output directly from the source condition. ControlNet [42], T2I-Adapter [43], and Uni-ControlNet [44] inject features extracted from the condition via dedicated encoder modules into a frozen diffusion backbone. Palette [45] and InstructPix2Pix [46] instead train end-to-end, concatenating the source image with the noisy latent. SDEdit [47] and Plug-and-Play Diffusion Features [48] manipulate a pre-trained model's sampling process without training, while BBDM [49] reformulates the diffusion trajectory as a stochastic bridge between domains. We adopt full-image generation because the indentation artifacts cannot be separated at pixel level, making mask-based inpainting fundamentally unsuitable.

3 Method

Fig. 2 illustrates the overall pipeline of DeTac, whose (a) dual-branch architecture and (b) joint optimization are introduced in Sec. 3.1 and Sec. 3.2, respectively. The hardware-aligned data collection process of the real-world paired dataset, on which DeTac is trained, is detailed in Sec. 3.3 and Fig. 3.

3.1 Concurrent Dual-Branch Prediction

Given a mixture image $\mathbf{I}_{\text{mix}} \in \mathbb{R}^{H\times W\times 3}$ captured inside a marker-based visual-tactile sensor (the image is RGB as the structured-light illumination is color-coded), our goal is to decouple the coupled tactile signals into a marker-free shading image $\hat{\mathbf{I}}_{\text{sha}} \in \mathbb{R}^{H\times W\times 3}$ for surface geometry reconstruction and a marker image $\hat{\mathbf{I}}_{\text{mar}} \in [0,1]^{H\times W}$ for marker tracking and force estimation. $\hat{\mathbf{I}}_{\text{mar}}$ is a soft mask image to account for defocus blur and ambiguous marker boundaries near the image periphery.

Previous methods rely on serial pipelines [21], where a marker mask is estimated and used for mask-based inpainting to obtain the shading image, therefore suffering from the aforementioned mask-sizing dilemma and error retention. To address these limitations, DeTac adopts a concurrent dual-branch architecture: The *shading restoration branch* performs conditional full-image generation to recover the shading image, while the *marker extraction branch* directly predicts the marker image.

Shading Restoration Branch. The shading restoration branch recovers the marker-free shading image $\hat{\mathbf{I}}_{\text{sha}}$ by restoring marker-occluded regions and suppressing marker-induced indentation artifacts in the mixed input $\mathbf{I}_{\text{mix}}$. We formulate this task as conditional full-image generation with a denoising diffusion model [18], allowing the model to exploit all available information in $\mathbf{I}_{\text{mix}}$ rather than relying on a predefined inpainting mask and discarding the marker-affected pixels as in mask-guided inpainting methods [21]. Specifically, our U-Net denoiser D_θ [50] takes the noisy target $\boldsymbol{x}_t$, the timestep t, and the mixed image $\mathbf{I}_{\text{mix}}$ as input at each discrete diffusion timestep $t \in \{1,\dots,T\}$:

$$\boldsymbol{\epsilon}_\theta = D_\theta(\boldsymbol{x}_t, t, \mathbf{I}_{\text{mix}}). \tag{1}$$

We use a predefined variance schedule $\{\beta_t\}_{t=1}^T$ with $0 < \beta_t < 1$, and define $\alpha_t = 1-\beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. During training, for a sampled timestep t, the noisy target $\boldsymbol{x}_t$ is obtained as

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{x}_0 = \mathbf{I}_{\text{sha}}$ denotes the GT noise-free target shading image, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ is Gaussian noise. The denoiser is optimized using the standard noise-prediction objective:

$$\mathcal{L}_{\text{sha}} = \mathbb{E}_{(\mathbf{I}_{\text{mix}},\mathbf{I}_{\text{sha}}),\boldsymbol{\epsilon},t}\left[\left\|\boldsymbol{\epsilon} - D_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{I}_{\text{sha}} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t, \mathbf{I}_{\text{mix}}\right)\right\|_2^2\right]. \tag{3}$$

To compute auxiliary losses (see Sec. 3.2) during training, the clean image is estimated from $\boldsymbol{\epsilon}_\theta$ as

$$\hat{\boldsymbol{x}}_0 = (\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta)/\sqrt{\bar{\alpha}_t}, \quad \hat{\mathbf{I}}_{\text{sha}} = \hat{\boldsymbol{x}}_0. \tag{4}$$

During inference, the model starts from pure Gaussian noise and iteratively samples $\boldsymbol{x}_{t-1}$ from $\boldsymbol{x}_t$:

$$\boldsymbol{x}_{t-1} = \left(\boldsymbol{x}_t - (1-\alpha_t)\boldsymbol{\epsilon}_\theta/\sqrt{1-\bar{\alpha}_t}\right)/\sqrt{\alpha_t} + \sigma_t\boldsymbol{z}, \qquad t = T,\dots,1, \tag{5}$$

where $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$. Following DDPM [18], we set $\sigma_t^2 = (1-\bar{\alpha}_{t-1})\beta_t/(1-\bar{\alpha}_t)$ with $\bar{\alpha}_0 = 1$. The restored shading image is given by the final denoised sample $\hat{\mathbf{I}}_{\text{sha}} = \hat{\boldsymbol{x}}_0$.

Marker Extraction Branch. The marker extraction branch isolates marker regions from the same mixed input $\mathbf{I}_{\text{mix}}$. The extracted marker image encodes the spatial distribution of markers under contact-induced deformation, from which marker-level quantities such as centroid positions can be obtained through post-processing. These quantities can then be used by downstream force-estimation algorithms [11], for example, by comparing displaced marker positions with their reference positions under the no-contact state. Specifically, we employ a lightweight fully convolutional network M_ϕ [51] to predict a soft marker image in a deterministic forward pass: $\hat{\mathbf{I}}_{\text{mar}} = M_\phi(\mathbf{I}_{\text{mix}})$.

During training, given the pseudo GT marker image $\mathbf{I}_{\text{mar}}$, whose acquisition will be detailed in Sec. 3.3, this branch is supervised using a pixel-wise Binary Cross-Entropy loss:

$$\mathcal{L}_{\text{mar}} = \text{BCE}(\hat{\mathbf{I}}_{\text{mar}}, \mathbf{I}_{\text{mar}}). \tag{6}$$

3.2 Cross-Branch Synergistic Optimization

Although the shading branch and the marker branch predict different targets, both are governed by the same marker distribution and its influence on the tactile image. We therefore introduce a cross-branch synergistic optimization that explicitly connects the two branches through marker-aware spatial structures. This optimization consists of two complementary objectives: a *surrounding*

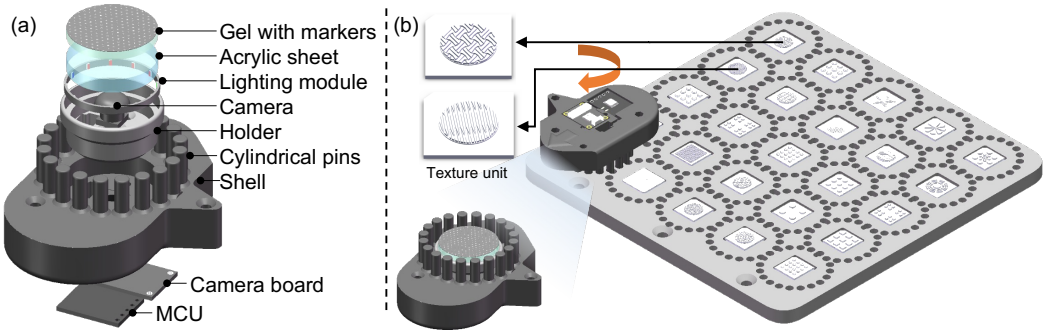


Figure 3: Hardware design and data acquisition setup. (a) Exploded view of the MC-Tac [52] sensor equipped with an outer shell featuring 20 cylindrical alignment pins arranged in a ring for repeatable positioning. (b) The data acquisition system features a collection pad with a 5×5 array of interchangeable texture units, each surrounded by 20 holes that mate with the sensor's alignment pins. Rotating the sensor and fully inserting the pins into different indexed holes under consistent force enable repeatable sampling of 20 contact poses per object before and after marker fabrication.

indentation loss that emphasizes restoration around marker neighborhoods, and a *change-marker consistency loss* that aligns marker prediction with restoration-induced appearance changes.

Surrounding Indentation Loss. In tactile images, marker-induced artifacts are not limited to marker-covered pixels. Due to local indentation introduced during marker fabrication, the neighborhood around each marker is often difficult to restore faithfully. To emphasize these critical regions, we construct an annular neighborhood around the predicted marker regions and apply focused supervision there. Specifically, let $\mathbf{B}_{\text{mar}} = \mathbb{I}(\hat{\mathbf{I}}_{\text{mar}} > \tau)$ denote the binarized marker prediction with threshold τ. The annular neighborhood around the predicted marker regions is defined as

$$\mathbf{R}_{\text{mar}} = \text{Dilate}(\mathbf{B}_{\text{mar}}, r) \setminus \mathbf{B}_{\text{mar}}, \tag{7}$$

where $\text{Dilate}(\cdot, r)$ denotes morphological dilation with radius r, and $\setminus$ denotes set difference. Based on this ring-shaped region, we define the surrounding indentation loss as

$$\mathcal{L}_{\text{sur}} = \|(\hat{\mathbf{I}}_{\text{sha}} - \mathbf{I}_{\text{sha}}) \odot \mathbf{R}_{\text{mar}}\|_1 / \left(\|\mathbf{R}_{\text{mar}}\|_1 + \delta\right), \tag{8}$$

where $\odot$ denotes element-wise multiplication, $\mathbf{R}_{\text{mar}}$ is broadcast to RGB channels, and δ is a small constant. This objective concentrates restoration supervision on marker neighborhoods, encouraging faithful reconstruction in the regions most affected by marker-induced indentation.

Change-Marker Consistency Loss. The pseudo GT marker images generated by SAM [53] with manual point annotations (see Sec. 3.3) may be inaccurate around marker boundaries, making $\mathcal{L}_{\text{mar}}$ alone insufficient for marker supervision. To provide complementary guidance, we introduce a change-marker consistency loss $\mathcal{L}_{\text{con}}$ that exploits the correlation between marker locations and restoration-induced appearance changes to provide mutual supervision.

We first compute a change map $\mathbf{I}_{\text{chg}} \in [0,1]^{H \times W}$ from the difference between the restored shading image and the mixed image. Specifically, we average the RGB difference across channels, sharpen the response with an exponent $\gamma = 2$, and normalize it spatially:

$$\bar{\mathbf{I}}_{\text{chg}} = \text{mean}_{\text{c}}(|\hat{\mathbf{I}}_{\text{sha}} - \mathbf{I}_{\text{mix}}|), \qquad \mathbf{I}_{\text{chg}} = \bar{\mathbf{I}}_{\text{chg}}^{\gamma} / (\text{max}_{\text{s}}(\bar{\mathbf{I}}_{\text{chg}}^{\gamma}) + \epsilon), \tag{9}$$

where mean_{c} denotes channel-wise averaging, max_{s} denotes the spatial maximum, and ϵ is a small constant. The change-marker consistency loss is then defined as

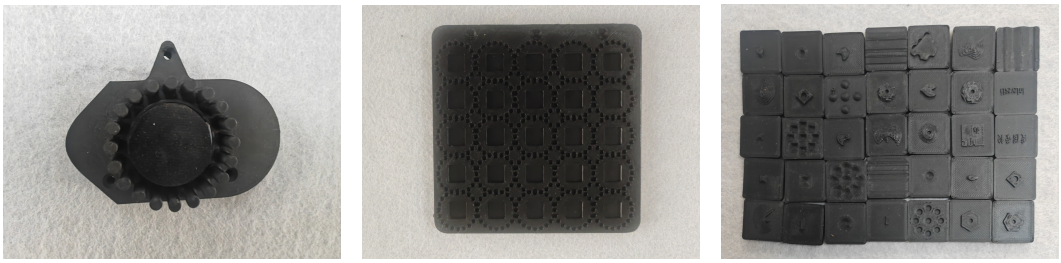
$$\mathcal{L}_{\text{con}} = \|\hat{\mathbf{I}}_{\text{mar}} - \mathbf{I}_{\text{chg}}\|_1. \tag{10}$$

Through this synergistic optimization, the two branches are connected through shared marker-aware spatial cues, improving both shading restoration fidelity and marker extraction accuracy.

Overall Objective. Together with the diffusion loss $\mathcal{L}_{\text{sha}}$ and the marker extraction loss $\mathcal{L}_{\text{mar}}$, the two auxiliary losses form the final training objective:

$$\mathcal{L} = \mathcal{L}_{\text{sha}} + \lambda_{\text{mar}}\mathcal{L}_{\text{mar}} + \lambda_{\text{sur}}\mathcal{L}_{\text{sur}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}, \tag{11}$$

where λ_{mar}, λ_{sur}, and λ_{con} are loss weights. In practice, $\mathcal{L}_{\text{sur}}$ and $\mathcal{L}_{\text{con}}$ are enabled after the first 100 epochs to avoid unstable auxiliary supervision at the early stage of training.



(a) Sensor with alignment pins (b) Collection pad (c) Texture unit samples

Figure 4: 3D printed hardware for dataset collection.

3.3 Real-World Paired Dataset Collection

Effectively training DeTac requires paired supervision for both shading restoration and marker extraction. Existing synthetic datasets [21] naively superimpose markers onto marker-free images, which ignores real marker-induced indentation artifacts and introduces a synthetic-to-real gap. To bridge this gap, we design a hardware-assisted pipeline to collect a real-world, well-aligned paired dataset. Specifically, we collect tactile data by repeatedly pressing the sensor against a diverse set of objects at multiple predetermined contact poses. Let $\mathbf{I}_{\text{mix}}^{o,k}$ denote the mixed tactile image acquired from object o at the k-th pose using the marker-based sensor, with corresponding marker-free shading image $\mathbf{I}_{\text{sha}}^{o,k}$ and marker image $\mathbf{I}_{\text{mar}}^{o,k}$. The resulting dataset is formulated as

$$\mathcal{D} = \left\{ \left(\mathbf{I}_{\text{mix}}^{o,k}, \mathbf{I}_{\text{sha}}^{o,k}, \mathbf{I}_{\text{mar}}^{o,k} \right) \;\middle|\; o \in \mathcal{O},\ k \in \{1, \ldots, K\} \right\}, \tag{12}$$

where $\mathcal{O}$ denotes the object set and K denotes the number of repeatable contact poses per object.

Hardware-Assisted Paired Data Acquisition. To obtain highly aligned shading and mixed tactile images, we divide data acquisition into two stages. First, we collect marker-free shading images $\mathbf{I}_{\text{sha}}$ using a pristine sensor elastomer. We then fabricate markers directly on this elastomer surface to reduce fabrication discrepancy and acquire corresponding mixed images $\mathbf{I}_{\text{mix}}$. Nevertheless, naively repeating the pressing process can still introduce pose drift. To ensure consistent contact conditions across the two acquisition stages, we design a mechanical alignment structure, as shown in Fig. 3. Assisted by this structure, we follow the same pose sequence before and after marker fabrication, enabling high-quality paired data acquisition of real physical contacts. We provide visualization of our 3D printed hardware for data collection, including the MC-Tac [52] sensor with alignment pins (Fig. 4a), the collection pad (Fig. 4b), and a subset of texture units (Fig. 4c).

Pseudo GT Marker Image Generation. After image acquisition, we generate pseudo GT marker images $\mathbf{I}_{\text{mar}}$ to supervise the marker extraction branch. To balance annotation efficiency and pixel-level accuracy, we use a prompt-based segmentation pipeline. We first manually annotate the center point of each marker in $\mathbf{I}_{\text{mix}}$. These point prompts, together with the original visual features, are then fed into SAM [53] to segment the marker and produce $\mathbf{I}_{\text{mar}}$. Compared with fully manual pixel-wise annotation, this strategy provides reliable spatial priors with substantially lower annotation cost.

Marker-Aware Augmentation for Cross-Sensor Generalization. To reduce overfitting to the fixed hardware specifications of our custom sensor and improve generalization across different marker-based tactile sensors, we introduce marker-aware spatial augmentations. Different sensor designs may exhibit different marker scales, densities, and boundary ambiguities due to variations in elastomer thickness and internal illumination. To simulate these variations, we apply aligned transformations to the triplet $(\mathbf{I}_{\text{mix}}, \mathbf{I}_{\text{sha}}, \mathbf{I}_{\text{mar}})$. Specifically, we use cropping followed by resizing to mimic changes in marker scale, and apply blurring guided by $\mathbf{I}_{\text{mar}}$ to simulate different levels of defocus. This strategy improves the cross-sensor robustness of DeTac.

4 Experiments

In this section, we conduct comprehensive experiments to evaluate DeTac. We first detail our experimental setup in Sec. 4.1, followed by comparisons with baselines in Sec. 4.2 and ablation studies in Sec. 4.3. Furthermore, we demonstrate how DeTac's decoupled outputs benefit downstream tasks in Sec. 4.4. The appendix provides more results.

Table 1: Quantitative results of shading restoration. The best results are highlighted in **bold**.

Method	Unseen pose				Unseen object			
	FID ↓	KID $\times 10^{-2}$ ↓	PSNR ↑	SSIM ↑	FID ↓	KID $\times 10^{-2}$ ↓	PSNR ↑	SSIM ↑
TELEA [23]	237.95	18.327	31.33	0.870	232.66	17.592	31.29	0.868
TacDiff [21]	158.94	11.447	31.52	0.866	155.63	10.654	31.40	0.863
LaMa [54]	358.16	37.535	29.09	0.804	360.27	37.117	29.16	0.803
SD 1.5 [20]	121.70	7.401	25.44	0.861	114.29	6.703	24.16	0.826
U-Net	108.35	7.489	38.56	0.950	111.28	7.750	39.03	0.942
Pix2Pix	112.53	7.472	38.53	0.951	109.66	7.065	38.98	0.948
dual-head network	104.27	6.923	37.99	0.949	105.18	6.794	38.37	0.947
TacDiff (enlarged mask)	172.01	11.898	30.93	0.852	180.96	12.384	30.84	0.847
LaMa (enlarged mask)	330.13	32.754	30.23	0.840	330.57	32.351	30.27	0.839
DeTac (w/o L_{sur})	25.78	0.733	34.64	0.946	22.59	0.698	35.05	0.945
DeTac (w/o L_{con})	22.65	0.689	37.77	0.949	19.44	0.634	38.28	0.948
DeTac (Full)	**20.44**	**0.683**	**39.45**	**0.953**	**17.06**	**0.573**	**39.07**	**0.950**

Table 2: Quantitative evaluation of marker extraction. The best result is highlighted in **bold**.

	UVTac [17]	SAM 3.1 [55]	DeTac (w/o $\mathcal{L}_{\text{sur}}$)	DeTac (w/o $\mathcal{L}_{\text{con}}$)	DeTac (Full)
MCLE ↓	0.21	0.41	0.00839	0.00832	**0.00762**

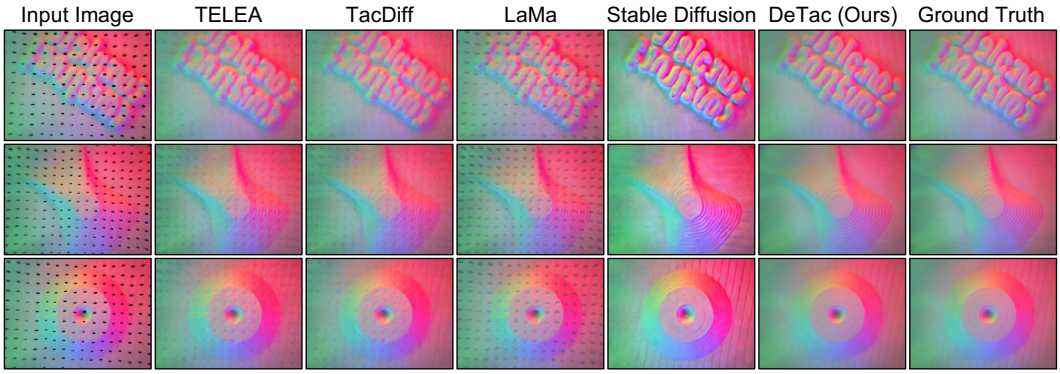


Figure 5: Qualitative comparison of shading restoration results.

4.1 Experimental Setup

Implementation Details. The whole pipeline is trained on a single NVIDIA RTX 4090 GPU for 300 epochs with a batch size of 2, which takes approximately 14 hours. The shading restoration and marker extraction branches are optimized with learning rates of 10^{-5} and 10^{-4} respectively. For the loss part, we use $\lambda_{\text{mar}} = 0.5$, $\lambda_{\text{sur}} = 0.05$, and $\lambda_{\text{con}} = 0.05$.

Baselines. For shading restoration, we compare DeTac against: the fast-marching-based TELEA [23]; two masked inpainting models, LaMa [54] and the marker removal model TacDiff [21], both trained on synthetic data generated from our dataset following the latter's methodology; Stable Diffusion (SD) 1.5 [20] fine-tuned on our data for full-image prediction; and three deterministic supervised baselines trained on the same paired dataset: a U-Net regression model, Pix2Pix, and a multi-task dual-head network. Our TacDiff reproduction strictly follows its original training protocol, using synthetically superimposed markers. We further evaluate enlarged-mask inpainting as an alternative strategy by dilating the marker masks to cover the surrounding indentation artifacts and applying them to both TacDiff and LaMa. For marker extraction, we evaluate DeTac against UVTac [17] and the segmentation model SAM 3.1 [55].

Metrics. To evaluate the quality of the restored shading images, we adopt Fréchet Inception Distance (FID) [56] and Kernel Inception Distance (KID) [57] to measure distributional similarity with the GT shading images in a deep feature space, as well as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [58] for pixel-level fidelity and structural similarity. For marker extraction, we introduce the Marker Centroid Localization Error (MCLE). For each marker,

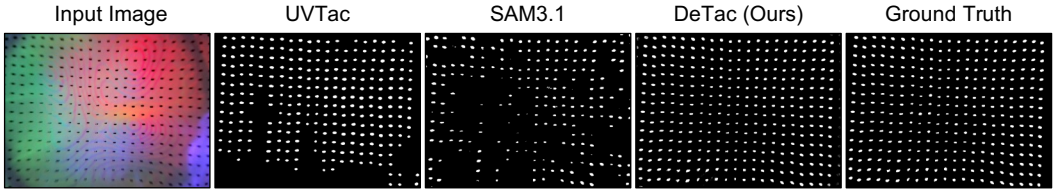


Figure 6: Qualitative comparison of marker extraction results.

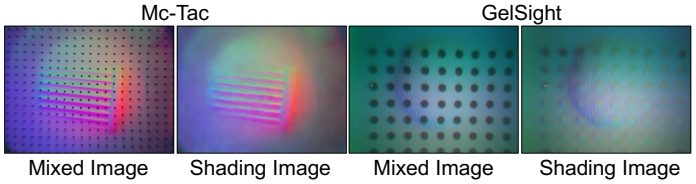


Figure 7: Generalization results of DeTac across different internal illuminations and sensor types.

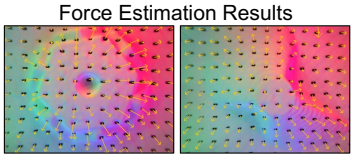


Figure 8: Force estimation results using extracted markers.

the predicted centroid $\hat{C}_i$ is computed as the average positions weighted by $\hat{\mathbf{I}}_{\text{mar}}$ within a circular region around the marker. MCLE is then defined as the L_2 distance in mm between $\hat{C}_i$ and the GT centroid C_i averaged over all markers, with lower values indicating more precise marker extraction.

Dataset. We set the number of poses per object to $K = 20$ and the number of objects to $|\mathcal{O}| = 54$, leading to 1080 pairs of tactile samples in total. We reserve 5 objects as "Unseen objects", and 1 pose from each of the other 49 objects as "Unseen poses" for testing. The rest are used for training.

4.2 Evaluation Results

Shading Restoration. As shown in Tab. 1, TELEA [23], TacDiff [21], and LaMa [54] exhibit poor performance across all metrics. Stable Diffusion [20] achieves lower FID and KID but suffers from degraded PSNR and SSIM. The deterministic supervised baselines perform worse in perceptual and distributional similarity, indicating that DeTac better restores realistic tactile appearance and fine details. Enlarged-mask TacDiff and LaMa still achieve lower performance than DeTac. DeTac achieves the best performance across all metrics. Qualitatively, as shown in Fig. 5, TELEA [23], TacDiff [21], and LaMa [54] leave visible residual marker traces, while Stable Diffusion [20] introduces unnatural stripe artifacts that distort the shading geometry. Our method produces the most natural results, closest to the GT.

Marker Extraction. As shown in Tab. 2, UVTac [17] and SAM 3.1 [55] exhibit significantly higher error compared to DeTac. Fig. 6 also shows that UVTac [17] and SAM 3.1 [55] produce coarse and incomplete masks, whereas DeTac successfully extracts all marker points with precise boundaries.

Cross-sensor Generalization. We evaluate generalization of DeTac on 1) the same type of sensor with different internal illumination, and 2) a different type of sensor [9]. As shown in Fig. 7, DeTac restores high-quality shading in both settings, demonstrating cross-sensor generalization.

4.3 Ablation Study

We ablate each loss term in Sec. 3.2 by removing them individually. As reported in Tab. 1 and Tab. 2, removing $\mathcal{L}_{\text{sur}}$ causes the most severe degradation in shading quality on both novel poses and novel objects, confirming that explicit supervision on marker-neighboring regions is critical. Removing $\mathcal{L}_{\text{con}}$ also degrades shading and slightly increases marker extraction error, indicating that the change map provides indispensable guidance for marker extraction. The full model achieves the best performance across all metrics, demonstrating that the two loss terms work synergistically.

4.4 Downstream Tasks

We demonstrate how DeTac's decoupled outputs benefit downstream tasks.

Force Estimation. Results in Fig. 8 demonstrate that our marker extraction enables precise displacement tracking for force estimation. We further quantitatively evaluate force estimation. The tactile sensor is mounted on a robot end-effector equipped with a six-axis force sensor. The measured forces are transformed from the force-sensor coordinate frame into the tactile-sensor coordinate frame and used as ground truth. We separately train a force-estimation network for each

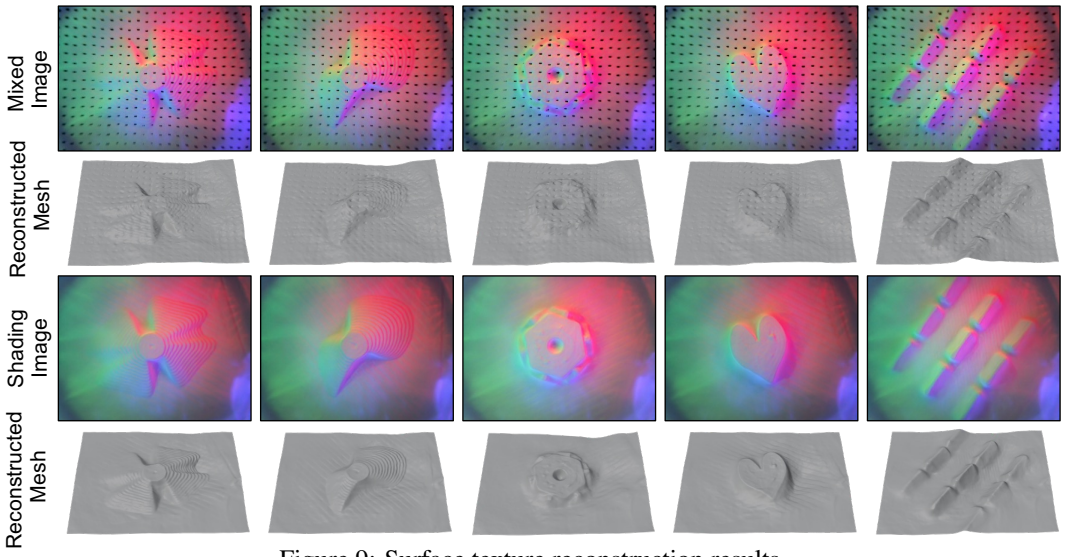


Figure 9: Surface texture reconstruction results.

Table 3: Force estimation results.

	Mixed	UVTac	SAM 3.1	DeTac
Mean force MAE (N) ↓	0.202	0.119	0.482	**0.089**

Table 4: Surface reconstruction results.

	Mixed	Shading
Chamfer Dist. (mm) ↓	0.165	**0.113**
RMSE (mm) ↓	0.111	**0.071**

method. The original mixed images and marker representations extracted by UVTac, SAM 3.1, and DeTac are fed into their respective networks and evaluated on the test set. As shown in Tab. 3, DeTac achieves the lowest mean force MAE, enabling more accurate force estimation than other methods.

Surface Reconstruction. As shown in Fig. 9, photometric stereo reconstruction [59] using our restored shading image yields significantly more accurate surface meshes compared to using the mixed tactile image, where residual marker artifacts introduce obvious geometric distortions. We further quantitatively evaluate surface reconstruction. The corresponding contact region extracted from the object's 3D model is used as the ground truth for evaluating the reconstructed surface. Each reconstructed mesh is aligned with the ground truth before computing the Chamfer distance and RMSE. As shown in Tab. 4, reconstruction using DeTac shading images achieves lower reconstruction metrics than using mixed images, demonstrating improved surface reconstruction accuracy.

5 Conclusion

We presented DeTac, a dual-branch full-image generation framework that concurrently recovers shading and marker components from a mixed tactile observation. A cross-branch synergistic optimization mechanism couples both branches for mutual supervision, while a hardware-assisted alignment structure provides paired real-world data. Experiments confirm DeTac's state-of-the-art decoupling accuracy, strong downstream performance, and effective cross-sensor generalization.

Limitations. First, the shading branch requires hundreds of sequential sampling steps at inference, limiting real-time deployment on latency-sensitive manipulation tasks; promising remedies include accelerated samplers or distillation-based compression. Second, our data acquisition pipeline is irreversible: markers are fabricated directly onto the elastomer and cannot be removed, so scaling the dataset requires manufacturing new sensors, which incurs additional hardware and labor overhead. Third, the current data collection setup does not support controlled shear forces, limiting the training data distribution.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant No. 52505029), Science and Technology Commission of Shanghai Municipality (Grant Nos. 25ZR1401191 and 24511103400), and Beijing High Innovation Plan (Contract No. 202604841443).

References

[1] M. K. Johnson and E. H. Adelson. Retrographic sensing for the measurement of surface texture and shape. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[2] C. Zhang, P. Hao, X. Cao, X. Hao, S. Cui, and S. Wang. VTLA: Vision-tactile-language-action model with preference learning for insertion manipulation. *Biomimetic Intelligence and Robotics (BIROB)*, page 100333, 2026.

[3] J. Huang, S. Wang, F. Lin, Y. Hu, C. Wen, and Y. Gao. Tactile-VLA: Unlocking vision-language-action model's physical knowledge for tactile generalization. *arXiv preprint arXiv:2507.09160*, 2025.

[4] N. Sunil, M. Tippur, A. Saumell, E. Adelson, and A. Rodriguez. Reactive in-air clothing manipulation with confidence-aware dense correspondence and visuotactile affordance. In *Proc. of Conference on Robot Learning (CoRL)*, 2025.

[5] M. Wu, L. Yang, J. Liu, W. Huang, L. Wu, Z. Chen, D. Ma, and H. Dong. UniTac2Pose: A unified approach learned in simulation for category-level visuotactile in-hand pose estimation. In *Proc. of Conference on Robot Learning (CoRL)*, 2025.

[6] B. Huang, J. Xu, I. Akinola, W. Yang, B. Sundaralingam, R. O'Flaherty, D. Fox, X. Wang, A. Mousavian, Y. Chao, et al. VT-refine: Learning bimanual assembly with visuo-tactile feedback via simulation fine-tuning. In *Proc. of Conference on Robot Learning (CoRL)*, 2025.

[7] J. D. Ferrandis, J. Moura, and S. Vijayakumar. Learning visuotactile estimation and control for non-prehensile manipulation under occlusions. In *Proc. of Conference on Robot Learning (CoRL)*, 2024.

[8] B. Huang, Y. Wang, X. Yang, Y. Luo, and Y. Li. 3D-ViTac: Learning fine-grained manipulation with visuo-tactile sensing. In *Proc. of Conference on Robot Learning (CoRL)*, 2024.

[9] W. Yuan, S. Dong, and E. H. Adelson. GelSight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017.

[10] E. Donlon, S. Dong, M. Liu, J. Li, E. H. Adelson, and A. Rodriguez. GelSlim: A high-resolution, compact, robust, and calibrated tactile- sensing finger. In *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[11] W. Yuan, R. Li, M. A. Srinivasan, and E. H. Adelson. Measurement of shear and slip with a GelSight tactile sensor. In *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

[12] D. Ma, E. Donlon, S. Dong, and A. Rodriguez. Dense tactile force estimation using GelSlim and inverse FEM. In *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[13] S. Wang, Y. She, B. Romero, and E. H. Adelson. GelSight Wedge: Measuring high-resolution 3d contact geometry with a compact robot finger. In *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[14] 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 *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, 2022.

[15] O. Tijani, Z. Chen, J. Deng, and S. Luo. MagicSkin: Balancing marker and markerless modes in vision-based tactile sensors with a translucent skin. *arXiv preprint arXiv:2512.06829*, 2025.

[16] Z. Hu, Z. Xu, S. Athar, J. P. Wachs, and Y. She. MuxGel: Simultaneous dual-modal visuo-tactile sensing via spatially multiplexing and deep reconstruction. *arXiv preprint arXiv:2603.09761*, 2026.

[17] W. Kim, W. D. Kim, J. Kim, C. Kim, and J. Kim. UVtac: Switchable UV marker-based tactile sensing finger for effective force estimation and object localization. *IEEE Robotics and Automation Letters (RA-L)*, 7(3):6036–6043, 2022.

[18] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[19] P. Dhariwal and A. Q. Nichol. Diffusion models beat GANs on image synthesis. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[20] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[21] N. Ou, Z. Chen, and S. Luo. Marker or markerless? mode-switchable optical tactile sensing for diverse robot tasks. *IEEE Robotics and Automation Letters (RA-L)*, 9(10):8563–8570, 2024.

[22] W. Huang, D. Lu, X. Zhong, Y. Tian, and Z. Wu. MVTS: multimodal visual-tactile sensor using a single camera. In *Proc. of IEEE International Conference on Systems, Man, and Cybernetics (SME)*, 2025.

[23] A. C. Telea. An image inpainting technique based on the fast marching method. *Journal of Graphics, GPU, and Game Tools (JGT)*, 9(1):23–34, 2004.

[24] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[25] 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*, 5(2):216–227, 2018.

[26] G. Liu, F. A. Reda, K. J. Shih, T. Wang, A. Tao, and B. Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proc. of European Conference on Computer Vision (ECCV)*, 2018.

[27] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[28] 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 (RA-L)*, 7(3):6187–6194, 2022.

[29] X. Jing, K. Qian, T. Jianu, and S. Luo. Unsupervised adversarial domain adaptation for sim-to-real transfer of tactile images. *IEEE Transactions on Instrumentation and Measurement (TIM)*, 72:1–11, 2023.

[30] M. Lambeta, P. Chou, S. Tian, B. H. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, D. Jayaraman, and R. Calandra. DIGIT: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters (RA-L)*, 5(3):3838–3845, 2020.

[31] M. K. Johnson, F. Cole, A. Raj, and E. H. Adelson. Microgeometry capture using an elastomeric sensor. *ACM Transactions on Graphics (TOG)*, 30(4):46, 2011.

[32] A. Padmanabha, F. Ebert, S. Tian, R. Calandra, C. Finn, and S. Levine. OmniTact: A multi-directional high-resolution touch sensor. In *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[33] N. Kuppuswamy, A. Alspach, A. Uttamchandani, S. Creasey, T. Ikeda, and R. Tedrake. Soft-bubble grippers for robust and perceptive manipulation. In *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[34] S. Dong, W. Yuan, and E. H. Adelson. Improved GelSight tactile sensor for measuring geometry and slip. In *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[35] C. Chorley, C. Melhuish, T. Pipe, and J. Rossiter. Development of a tactile sensor based on biologically inspired edge encoding. In *Proc. of International Conference on Advanced Robotics (ICAR)*, 2009.

[36] B. Winstone, C. Melhuish, T. Pipe, M. Callaway, and S. Dogramadzi. Toward bio-inspired tactile sensing capsule endoscopy for detection of submucosal tumors. *IEEE Sensors Journal*, 17(3):848–857, 2016.

[37] B. Ward-Cherrier, L. Cramphorn, and N. F. Lepora. Tactile manipulation with a TacThumb integrated on the open-hand M2 gripper. *IEEE Robotics and Automation Letters (RA-L)*, 1(1): 169–175, 2016.

[38] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[39] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. of IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

[40] L. Stacchio. Train stable diffusion for inpainting, 2023. https://github.com/lorenzo-stacchio/Stable-Diffusion-Inpaint. Accessed: 2026-05-28.

[41] X. Ju, X. Liu, X. Wang, Y. Bian, Y. Shan, and Q. Xu. BrushNet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *Proc. of European Conference on Computer Vision (ECCV)*, 2024.

[42] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *Proc. of IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[43] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proc. of Association for the Advancement of Artificial Intelligence (AAAI)*, 2024.

[44] S. Zhao, D. Chen, Y. Chen, J. Bao, S. Hao, L. Yuan, and K. K. Wong. Uni-ControlNet: All-in-one control to text-to-image diffusion models. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[45] C. Saharia, W. Chan, H. Chang, C. A. Lee, J. Ho, T. Salimans, D. J. Fleet, and M. Norouzi. Palette: Image-to-image diffusion models. In *Proc. of the ACM SIGGRAPH Conference and Exhibition On Computer Graphics and Interactive Techniques (SIGGRAPH)*, 2022.

[46] T. Brooks, A. Holynski, and A. A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[47] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J. Zhu, and S. Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *Proc. of International Conference on Learning Representations (ICLR)*, 2022.

[48] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[49] B. Li, K. Xue, B. Liu, and Y. Lai. BBDM: image-to-image translation with brownian bridge diffusion models. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[50] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proc. of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[51] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[52] J. Ren, J. Zou, and G. Gu. MC-Tac: Modular camera-based tactile sensor for robot gripper. In *Proc. of International Conference on Intelligent Robotics and Applications (ICIRA)*, 2023.

[53] N. Ravi, V. Gabeur, Y. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, et al. SAM 2: Segment anything in images and videos. In *Proc. of International Conference on Learning Representations (ICLR)*, 2025.

[54] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proc. of IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022.

[55] N. Carion, L. Gustafson, Y. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, et al. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

[56] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[57] M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying MMD GANs. In *Proc. of International Conference on Learning Representations (ICLR)*, 2018.

[58] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)*, 13(4): 600–612, 2004.

[59] H.-J. Huang. Gelsight sdk, 2024. https://github.com/joehjhuang/gs_sdk. Accessed: 2026-05-28.

[60] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[61] P. Ramachandran, B. Zoph, and Q. V. Le. Searching for activation functions. In *Proc. of International Conference on Learning Representations (ICLR)*, 2018.

[62] Y. Wu and K. He. Group normalization. In *Proc. of European Conference on Computer Vision (ECCV)*, 2018.

[63] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proc. of International Conference on Machine Learning (ICML)*, 2010.

[64] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *Proc. of International Conference on Learning Representations (ICLR)*, 2019.

[65] Realman. RM75-6F: 7-DOF Robotic Arm 6-Axis-Force Version, 2024. https://www.realman-robotics.com/rm75-6f.html. Accessed: 2026-06-04.

DeTac: Dual-branch Decoupling of Texture and Force Signals in Vision-Based Tactile Sensors

Appendix

In this appendix, we provide additional implementation details in Sec. A, including a detailed description of the network architecture and training configurations. We further present a visualization of the paired dataset in Sec. B, along with additional experiments on a real-world downstream fabric classification task in Sec. C.

A Additional Implementation Details

A.1 Network Architecture Specifications

Shading Restoration Branch. The shading restoration network D_θ adopts a U-Net architecture with 22 residual blocks [60]. It directly processes a 6-channel input — formed by the pixel-space concatenation of the mixed image $\mathbf{I}_{\text{mix}}$ and the noisy target $\boldsymbol{x}_t$ — to output a 3-channel noise prediction. Specifically, the encoder features 4 downsampling stages with block output channels of (64, 128, 256, 256), where each stage contains 2 residual blocks. The bottleneck consists of 2 residual blocks. The decoder features 4 corresponding upsampling stages; to effectively fuse the concatenated skip connections, each decoder stage employs 3 residual blocks. Every residual block utilizes 2 convolutional layers paired with Swish activations [61] and Group Normalization [62]. Finally, a terminal convolutional layer projects the features to the target 3-channel noise space.

Marker Extraction Branch. The marker extraction network M_ϕ is constructed as a lightweight fully convolutional network [51] comprising three layers. The channel dimensions progress as $3 \to 32 \to 64 \to 1$, with each convolution using a kernel size of 3, a stride of 1, and a padding of 1. ReLU activations [63] follow the first two convolutions, while the final convolution outputs logits without an activation function. The soft marker image is obtained by applying a sigmoid function to these logits. During training, `BCEWithLogitsLoss` directly operates on the logits for numerical stability.

A.2 Additional Training Details

For the denoising diffusion forward process, the total number of discrete training timesteps is set to $T = 1000$. The morphological dilation radius used to define the annular neighborhood $\mathbf{R}_{\text{mar}}$ for the surrounding indentation loss $\mathcal{L}_{\text{sur}}$ is set to $r = 2$ pixels. Optimization is performed using the AdamW optimizer [64] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and weight decay $= 0.01$.

A.3 Data Collection Details

During paired-data collection, 20 alignment pins fix pose and press depth, while a 200 g weight maintains consistent press force. To evaluate the effect of press-force variation, images captured with 200 g and 500 g weights at identical depths yield a per-pixel MAE of 3.072 (0–255 scale), indicating a constant 200 g force does not significantly limit dataset visual diversity. Markers are directly embedded in the same elastomer to minimize fabrication-induced pattern shifts. In Fig. A1, the left shading image, middle mixed image, and right image (their pixel-wise difference) reveal indentation artifacts around markers.



Figure A1: Pixel-wise errors caused by indentation artifacts.

B Paired Dataset

We further visualize in Fig. A2 one paired data sample from each of the 54 texture units, resulting in 54 paired data examples in total.

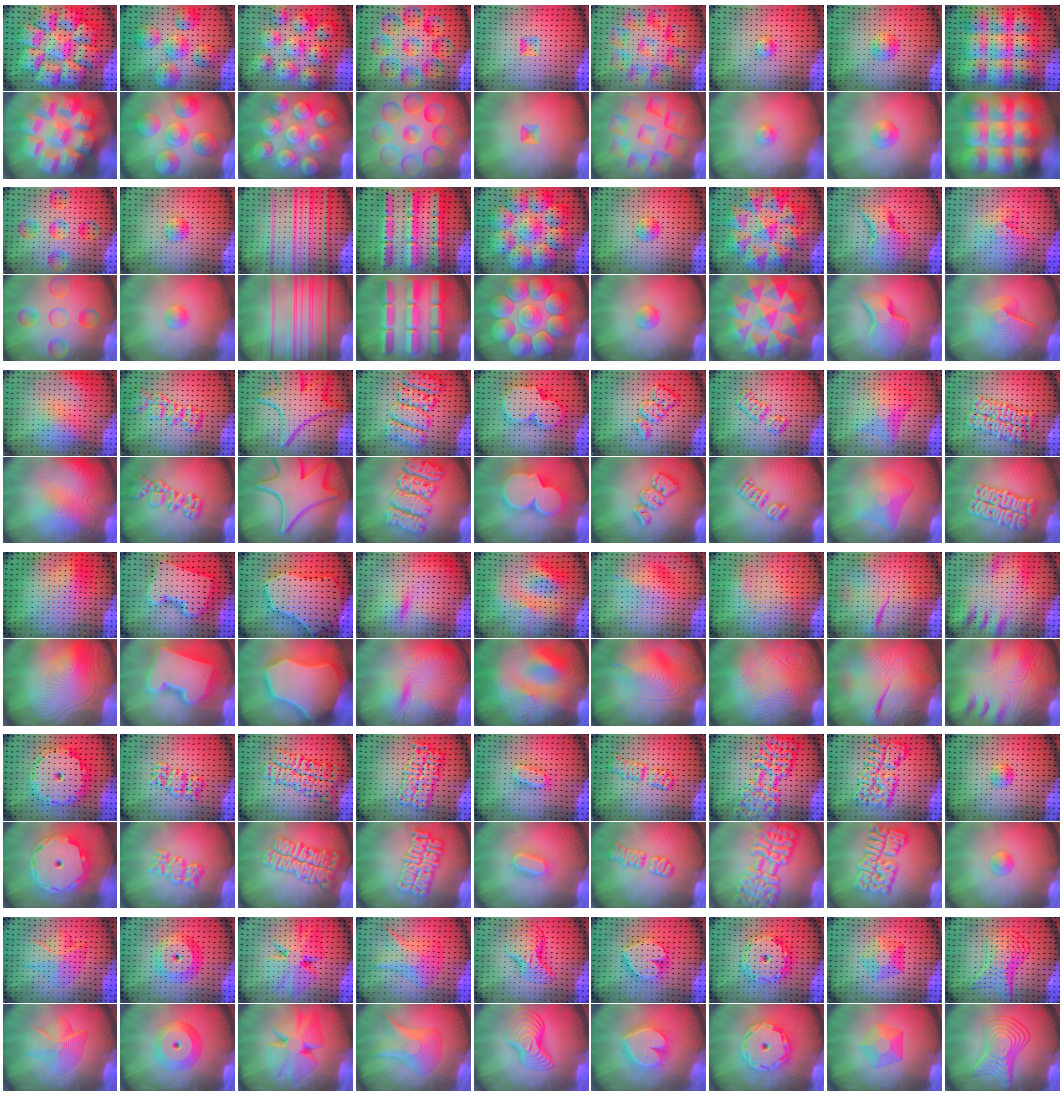


Figure A2: Visualization of paired data samples from each of the 54 texture units, arranged in 12 rows and 9 columns. Odd rows display the mixture image $\mathbf{I}_{\mathrm{mix}}$ with markers, and even rows show the corresponding shading image $\mathbf{I}_{\mathrm{sha}}$ directly below each mixture image.

C Additional Experiments

To further verify that our model successfully decouples the signal, we conduct a real-world downstream fabric classification experiment to assess whether the separated shading image removes marker-induced interference while preserving sufficient physical information for practical tasks. We first train a ResNet-based [60] texture classification network on data collected from a markerless tactile sensor. The sensor is mounted on a RealMan RM-75 robotic arm [65], which presses against a fabric sample to perceive its surface texture and classify it among several fabric categories.

We then replace the sensor with a marker-based sensor that is largely aligned with the markerless one, differing only in slight hardware variations. Using the same pre-trained classification network, we compare two evaluation settings: (1) directly feeding the raw mixed image into the classifier, and

Table A1: Fabric texture classification success rates across three fabric types.

Input Type	Success Rate			Overall
	Fabric A	Fabric B	Fabric C	
Raw mixed image	16.7%	43.3%	90.0%	50.0%
Our marker-removed shading image	**83.3**%	**73.3**%	**96.7**%	**84.4**%

(2) feeding the marker-removed shading image predicted by DeTac into the classifier. For each of the three fabric types, we conduct 30 trials. The success rates are reported in Tab. A1. The results demonstrate that our method effectively eliminates marker interference, and the resulting shading images can be directly leveraged for downstream tasks. Although the robot arm carries a wrist-mounted RealSense D405C camera, its data is disregarded during experiment, our classification decisions are based exclusively on tactile signals.