IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Tacser and Action-Conditioned Latent Filter for Generalizable Robotic Surface Perception

Anirvan Dutta, Yerkebulan Massalim, Etienne Burdet, and Mohsen Kaboli

***Abstract*—Perceiving the physical properties of different surfaces/textures via tactile sensing has been a long-standing problem in robotics. Most prior work has been limited to discriminative models that classify textures into a fixed set of categories. However, to enable seamless and efficient autonomous manipulation, robots must infer physical properties as structured, continuous variables rather than as discrete class labels. In this work, we present a novel deep state-space model (DSSM) to learn and infer key causal textural properties in an unsupervised manner. Using variational inference to solve the DSSM, our proposed *Latent Filter* allows robotic systems to perceive textures in a continuous and generalizable manner. In addition, we explore a novel interaction approach: *Tacser* (Tactile Enhancer), to further enhance tactile sensing through vibrations induced by high-frequency micro-movements and thereby improve perception. We evaluated our approach against state-of-the-art techniques and performed extensive ablation studies to demonstrate its effectiveness. This work advances tactile-based texture perception, providing a generalizable and comprehensive framework for robotics.**

***Index Terms*—Perception for Grasping and Manipulation; Probabilistic Inference; Interactive Perception; Tactile Sensing.**

I. Introduction

AUTONOMOUS robotic systems operating in dynamic environments must accurately perceive the physical properties of their surroundings and the objects within them, including shape, inertia, surface characteristics, and stiffness [1]. Among these, estimating surface properties remains particularly challenging, despite advances in tactile sensing and interactive exploration [2]–[4]. Surface properties, or textures, are inherently complex and difficult to characterize, encompassing frictional properties, micro/macrostructures (such as roughness), spatial periodicity, etc. Precise perception of these properties is essential for enabling robotic systems to perform sophisticated dynamic tasks, such as slip prevention during grasping [5], fine-grained manipulation [6], and object identification or recognition [7].

Received June 5 2025; accepted October 8 2025. Date of publication; date of current version. This letter was recommended for publication by Editor M. Vincze upon evaluation of the reviewers' comments. This work was supported in part by the BMW Group, in part by the EU H2020 INTUITIVE project under Grant 861166, and in part by the EU Horizon PHASTRAC project under Grant 101092096.

Anirvan Dutta is with the Imperial College of Science, Technology and Medicine, SW7 2AZ London, U.K, and was also with BMW Group AG, Munich during this work (e-mail: a.dutta22@imperial.ac.uk)

Yerkebulan Massalim is with Parrot Drones, Paris, France and was with Actronika SAS, Paris, France and Laboratoire des systèmes perceptifs (LSP), Ecole Normale Superieure (ENS), Paris, during the work.

Etienne Burdet is with the Imperial College of Science, Technology and Medicine, SW7 2AZ London, U.K.

Mohsen Kaboli is with the BMW Group AG, Parkring 19, 85748, Munich, Germany, and also with Eindhoven University of Technology (TU/e), 5612 AZ Eindhoven Netherlands.

Digital Object Identifier (DOI): see top of this page.

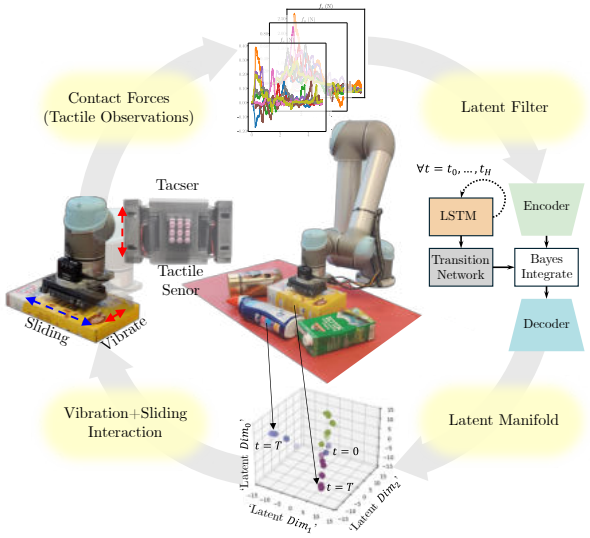

Fig. 1: Setup for robotic texture perception.

Researchers have focused primarily on developing robotic systems equipped with tactile sensing that recognize textures using hand-crafted or deep learning-based feature extraction to classify them using discriminative models [8]. Although effective in specific contexts, these approaches do not adequately capture the vast diversity of real-world textures or surface parameters. A promising alternative is regression-based models that estimate continuous textural properties. However, current state-of-the-art analytical models such as Coulomb or LuGre [9] focus only on the frictional aspect. Furthermore, previous studies employing sliding or pressing interactions (with camera-based tactile sensors) have largely overlooked the role of interaction parameters in the extraction of tactile features and the modeling of perceived interactive tactile data. This highlights the need for an action-conditioned, unsupervised and learnable regression model capable of inferring causal textural invariants—without reliance on ground truth properties—from complex tactile signals generated during surface (texture) interactions with a robot [10].

To address such a challenging problem, we take inspiration from humans, who can perceive and distinguish textures with remarkable precision and minimal interaction effort. Extensive research has explored to understand such human tactile perception [11], [12], demonstrating that people can accurately perceive the physical properties of textures and describe these sensations using adjectives such as "smooth," "rough," or "sticky" during exploration of surfaces with varying textural qualities along with their intensity (how smooth or rough) [13]. This suggests a subtle and sophisticated process of tactile-based interactive texture perception. Inspired by such perceptual capability, in this study we present a novel unsu-

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

pervised variational inference-based action-conditioned deep state-space modeling and learning scheme to capture the dynamics of the texture-robot interaction and infer textural properties through a *latent filter*.

Furthermore, the role of human skin in texture perception has been explored through theories such as the 'Duplex Theory' by Katz [14] and further refined by Hollins and Risner [15], which suggests that the multilayered structure of human skin and specialized mechanoreceptors allow advanced texture perception by integrating touch, temperature, and motion stimuli. Despite progress, artificial tactile sensors still lag behind human sensing in complexity and versatility, which is critical to achieving human-level perceptual capabilities [16]. A recent study [17] demonstrated that human texture perception improves with the induction of external vibration stimuli during exploration due to richer contact information. Inspired by this, we developed a novel device - *Tacser* which provides the capability to induce high-frequency vibration independently while performing sliding-based texture exploration. The overview of the proposed framework is illustrated in Figure 1, integrating *tacser* with the *latent filter* to advance tactile-based texture perception in robotics, ensuring both generalizability and comprehensiveness.

II. Related Work

One of the earliest works on artificial texture perception was proposed by Tada et al. [18], which demonstrated that just the variance in the outer strain gauge pressure obtained by an anthropomorphic finger during sliding could effectively discriminate among four different surfaces. Subsequent research expanded this concept to a broader range of texture recognition, utilizing various features in the tactile signals - vibrations, pressure variations, etc. The features ranged from raw sensor data [19], frequency-related information using Fourier [20], wavelet [21] [22] or statistical features such as in [23] and combination [24]. The capabilities of the features were significantly limited, primarily allowing for the discrimination of only a small set of textures. This limitation underscored the importance of analyzing key features in tactile-sensor interactions during texture perception.

Fishel et al. [25] addressed this issue by proposing statistical features that approximate the mechanical properties of roughness, traction, and fineness using a BioTac sensor and a large texture set. Chu et al. [26] built on this work by combining sensor-specific static and dynamic features. Kaboli et al. [27] introduced robust sensor-agnostic statistical features inspired by Hjorth parameters: activity, mobility, and complexity. Although these features were discriminative, they lacked physical alignment with mechanical properties, which complicated the generalization to novel textures [28].

Following this, numerous studies have explored deep learning techniques [29] to autonomously extract features from tactile data using different tactile sensing technologies [30] under varying interaction conditions. Prominent approaches involve computing spectrograms of time-varying tactile signals and leveraging CNNs [31]–[33] to focus on spatial and temporal patterns, often combined with multiple scales [34] or recurrent networks to capture the sequential information [35], [36] and incorporated complex attention mechanism [37]. Such discriminative deep learning techniques require extensive training data with human-annotated texture labels and are limited to predefined classes, limiting the generalizable capability. More recently, vision-based tactile sensors have enabled texture recognition by capturing only microstructural details from static pressing interactions [38]. However, these methods often overlook vibrational and directional cues, which are obtained only via high-frequency sensing and dynamic interactions. A large body of work has explored visuo-tactile fusion [39] or cross-modality [40]–[42] from vision to touch within the texture perception problem, which remains out of the scope of the current study.

Recent efforts toward generalization using domain adaptation [43], [44] and clustering [45] often require extensive tuning, underscoring the need for fully unsupervised learning approaches. While unsupervised methods such as PCA [46], [47], sparse representations [48], autoencoders [49], and variational autoencoders [40] have been explored, they often overlook the sequential and dynamic nature of tactile interactions. Incorporating sequential structure is critical for downstream manipulation tasks and overcomes the limitations of static, fixed-sequence models such as VAEs. Autoregressive models [50] have extended these efforts to temporal settings, but remain limited to next-step prediction without explicitly modeling interaction dynamics or action parameters. Consequently, extracting low-dimensional, physically grounded representations of texture and stiffness remains an open challenge. Deep state-space models (DSSMs) under variational inference show promise in capturing physically meaningful representations through unsupervised learning [51], [52]. However, their application to robotic texture perception problem, where interaction parameters are essential for action-conditioned inference, remains largely unexplored [53]. Moreover, inducing vibrations during dynamic interactions poses non-trivial engineering challenges, as it requires precise synchronization of motion and actuation. Unlike prior work [54], which applied vibration only in static contact scenarios, our objective was to allow vibration during sliding interactions to capture richer tactile signals. This study aims to bridge these gaps through the following contributions.

1) We propose an action-conditioned sequential deep state-space model (DSSM) that captures complex surface–robot interaction dynamics by structuring the latent space into directly and indirectly observable components. This facilitates the inference of causal and physically meaningful textural properties.
2) We introduce a novel unsupervised variational inference framework—*Latent Filter*—for learning the DSSM and estimate properties in a temporally coherent latent manifold. Our approach relaxes the requirement for complete observation at every time step, enabling effective inference from partial sequences and improving suitability for real-time robotic applications.
3) We develop *Tacser*, a mechanism that deliberately induces high-frequency micro-motions during dynamic interaction, thereby enhancing excitation of contact dynamics, improving texture perception.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

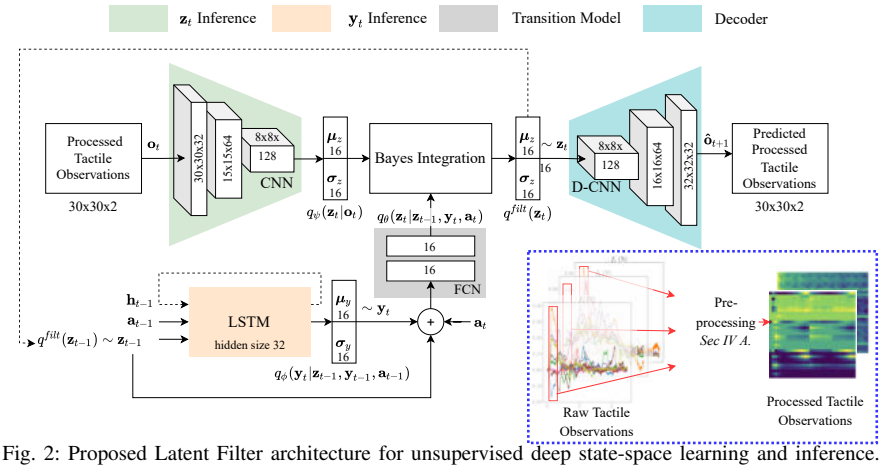

Fig. 2: Proposed Latent Filter architecture for unsupervised deep state-space learning and inference. Dotted black lines indicate recurrent connections carrying variables from $t-1$ to t. The blue, dotted inset shows the preprocessing pipeline.

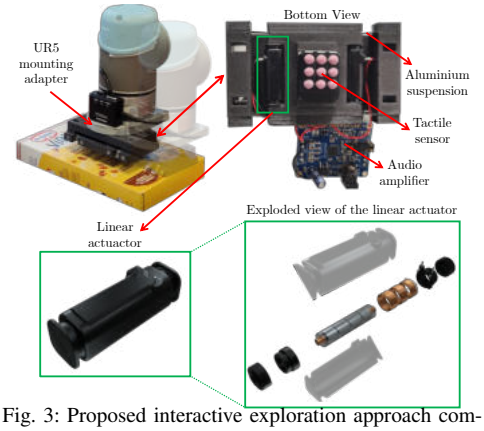

Fig. 3: Proposed interactive exploration approach comprising the tactile sensor and vibration mechanism (*Tacser*).

We validated the proposed approach through extensive real-robot experiments and comprehensive ablation studies on interaction, demonstrating consistent advantages over state-of-the-art baselines.

III. Proposed Method

A. Problem Definition

We model the interaction between the texture surface and the robot as a discrete nonlinear dynamical system with tactile observations (contact forces) $\mathbf{o}_{1:T} = (\mathbf{o}_1, \mathbf{o}_2, .., \mathbf{o}_T), \mathbf{o}_t \in \mathcal{O} \subset \mathbb{R}^{n_o}$ in discrete time steps $t = 1, .., T$ and actions $\mathbf{a}_{1:T} = (\mathbf{a}_1, \mathbf{a}_2, .., \mathbf{a}_T), \mathbf{a}_t \in \mathcal{A} \subset \mathbb{R}^{n_a}$.

B. Enhancing Tactile Perception via Induced Vibration

The characteristics of vibration propagation vary depending on the physical properties of the texture, which can be used to amplify the vibrotactile components to identify the textural properties. We introduce an interactive exploration system that incorporates a tactile sensor attached to a novel vibrating mechanism, as depicted in Figure 3, allowing independent vibration induction during dynamic textural object exploration.

1) Tactile Sensor

The tactile sensor used in this study is commercially available Contactile [55] sensor, which consists of tactile sensor arrays featuring silicone pillars arranged in a rectangular grid of 3×3. Each pillar has a diameter of $6\,mm$ and the pillars are spaced $7\,mm$ apart. This sensor can measure calibrated 3D displacement, 3D force, and vibration on each sensing element (referred to as a taxel), as well as global 3D force and global 3D torque using optical transduction. Furthermore, we found that the viscoelastic properties of the silicone pillars are highly effective in propagating vibrations to textured surfaces. The tactile sensor records normal and shear forces with a resolution of $\pm \leq 0.05N$ at a frequency of $1000\,Hz$, providing sufficient spatial and temporal resolution for the texture perception problem.

2) Vibration Mechanism-Tacser

Vibrations are induced using a pair of HapCoil linear actuators manufactured by Actronika [56]. The actuator has a resonant frequency of $70Hz$ and is capable of reaching an acceleration rate of 1-2g up to $1kHz$. Figure 3 shows an exploded and side view of the actuator. The moving part is suspended by a four-blade suspension system made of aluminum flexures with dimensions $10 \times 8 \times 0.1mm$. Due to the flexure system, the system bandwidth is flat in the range 100–$500Hz$ with $\pm 2dB$ ripples, coinciding with the response of a damped mass-spring system. A compact size of the system minimizes the non-linearities emanated from the 3D printed parts to obtain perfect reflection of the input signal as means of vibrations. We use an audio amplifier board to drive the actuators (MAX9744, Stereo 20W Class D Audio Amplifier, Adafruit). The input signal (see Figure 4) is transmitted to the audio amplifier via an audio jack affixed to the board. This novel design enables the robotic system to induce vibrations orthogonal to the lateral sliding interaction, thereby enhancing tactile sensing.

3) Interaction parameters

The interactive exploration action, represented as $\mathbf{a}_t = \{v_{x_t}, d_{y_t}, f_{z_t}\}$, consists of a combination of vibration and sliding parameters. The sliding was performed orthogonally to the vibration to help disentangle the spatial and vibration characteristics [57]. Formally formulated as a tuple of three components: *sliding velocity* v_{x_t}, *vibration amplitude* d_{y_t} and *normal contact force* f_{z_t}. The low-velocity sliding allows for the perception of spatial and low-frequency components of the texture, whereas the high-frequency components are perceived by the vibrations in the orthogonal direction. Both depend on the normal contact force exerted by the robotic system when interacting with the texture. To achieve varying frequencies of excitation, we designed a chirp signal to be input into the actuators. A chirp signal is a sinusoidal waveform whose frequency increases over time, typically represented in a general form:

$$d_y(t) = A(t)\,\sin\left[2\pi\left(f_0 t + \frac{k}{2}t^2\right)\right] \tag{1}$$

where t is time, A is the time-varying amplitude, f_0 the initial frequency, and k is the chirp rate, i.e., rate of frequency change. The lateral sliding velocity v_{x_t} is kept constant during a particular interaction but varied in different interactions ranging from 0.01 to 0.05 m/s. In addition, the parameterized

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

normal contact force f_{z_t} is varied between 0.5 to $1.5\,N$ which provides the most discriminate tactile observations [27].

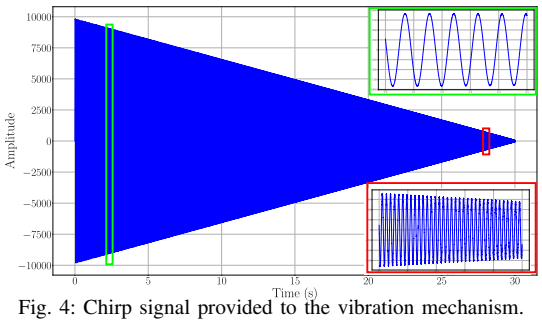

Fig. 4: Chirp signal provided to the vibration mechanism.

C. Deep State-Space Modeling: Latent Filter

A deep state-space model (DSSM) is employed to capture interaction dynamics by leveraging time-series and action-conditioned tactile data through low-dimensional latent variables $\mathbf{s}_{1:T} = (\mathbf{s}_1, \mathbf{s}_2, .., \mathbf{s}_T)$ with $\mathbf{s}_t \in \mathcal{S} \subset \mathbb{R}^{n_s}$ that represent the underlying state of the system. The objective is to model the joint probability $p(\mathbf{o}_{1:T}, \mathbf{s}_{1:T}|\mathbf{a}_{1:T})$, and perform variational inference by maximizing the likelihood of observations.

$$p(\mathbf{o}_{1:T}|\mathbf{a}_{1:T}) = \int p(\mathbf{o}_{1:T}, \mathbf{s}_{1:T}|\mathbf{a}_{1:T})\, d\mathbf{s}_{1:T} \tag{2}$$

We assume a generative model with an underlying latent dynamical system with

$$\begin{aligned} p(\mathbf{o}_{1:T}|\mathbf{a}_{1:T}) &= \int p(\mathbf{o}_{1:T}|\mathbf{s}_{1:T}, \mathbf{a}_{1:T})\, p(\mathbf{s}_{1:T}|\mathbf{a}_{1:T})\, d\mathbf{s}_{1:T} \\ &= \int \prod_{t=1}^{T} p(\mathbf{o}_t|\mathbf{o}_{1:t-1}, \mathbf{s}_t, \mathbf{s}_{1:t-1}, \mathbf{a}_{1:t}) \\ &\quad p(\mathbf{s}_t|\mathbf{o}_{1:t-1}, \mathbf{s}_{t-1}, \mathbf{s}_{1:t-2}, \mathbf{a}_t, \mathbf{a}_{1:t-1})\, d\mathbf{s}_{1:t} \end{aligned} \tag{3}$$

Eq. 3 becomes computationally very expensive with increasing time steps, as with each time step, the conditional variables increase. Therefore, we use Markov's assumption to simplify

$$\begin{aligned} p(\mathbf{o}_t|\cancel{\mathbf{o}_{1:t-1}}, \mathbf{s}_t, \cancel{\mathbf{s}_{1:t-1}}, \cancel{\mathbf{a}_{1:t}}) &= p(\mathbf{o}_t|\mathbf{s}_t) \\ p(\mathbf{s}_t|\cancel{\mathbf{o}_{1:t-1}}, \mathbf{s}_{t-1}, \cancel{\mathbf{s}_{1:t-2}}, \mathbf{a}_t, \cancel{\mathbf{a}_{1:t-1}}) &= p(\mathbf{s}_t|\mathbf{s}_{t-1}, \mathbf{a}_t) \end{aligned}$$

resulting in the following simplified generative model:

$$\begin{aligned} p(\mathbf{o}_{1:T}|\mathbf{a}_{1:T}) &= p(\mathbf{o}_1|\mathbf{s}_1, \mathbf{a}_1)p(\mathbf{s}_1) \\ &\quad \int \prod_{t=2}^{T} p(\mathbf{o}_t|\mathbf{s}_t)\, p(\mathbf{s}_t|\mathbf{s}_{t-1}, \mathbf{a}_t)\, d\mathbf{s}_t \end{aligned} \tag{4}$$

We hypothesize that modeling the dynamics of robot-texture interactions enables the extraction of causal physical factors underlying the process. To support this, we introduce a structural assumption in the latent space, inspired by observability in control theory [58]. Specifically, we distinguish between directly observable components, inferred from a single observation, and indirectly observable components, which require multiple time steps for accurate estimation. This partition facilitates analytical posterior computation while avoiding the cost of linearizing the observation model [51].

$$p(\mathbf{s}_t) = p(\mathbf{z}_t, \mathbf{y}_t) = p(\mathbf{z}_t|\mathbf{y}_t)\, p(\mathbf{y}_t) \tag{5}$$

with directly observable variable $\mathbf{z}_t \subset \mathbb{R}^{n_z}$ and indirectly observable part as $\mathbf{y}_t \subset \mathbb{R}^{n_y}$ with $n_s = n_z + n_y$, resulting in the following

$$\begin{aligned} p(\mathbf{o}_{1:T}|\mathbf{a}_{1:T}) = \iint \prod_{t=2}^{T} & p(\mathbf{o}_t|\mathbf{z}_t)\, p(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{y}_t, \mathbf{a}_t) \\ & p(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathbf{a}_{t-1})\, d\mathbf{y}_{t-1} d\mathbf{z}_{t-1} \end{aligned} \tag{6}$$

To compute the observation likelihood, we introduce variational distribution $q_\theta(\mathbf{z}_{1:T}, \mathbf{y}_{1:T}) \sim p(\mathbf{z}_{1:T}, \mathbf{y}_{1:T}|\mathbf{o}_{1:T}, \mathbf{a}_{1:T})$. The Evidence Lower Bound Objective (ELBO) for the generative model in Eq. 6 is formulated from the KL Divergence inequality:

$$\begin{aligned} D_{\text{KL}} = -\iint & q_\theta(\mathbf{z}_{1:T}, \mathbf{y}_{1:T}|\mathbf{o}_{1:T}, \mathbf{a}_{1:T}) \\ & \log\left[\frac{p(\mathbf{z}_{1:T}, \mathbf{y}_{1:T}|\mathbf{o}_{1:T}, \mathbf{a}_{1:T})}{q_\theta(\mathbf{z}_{1:T}, \mathbf{y}_{1:T}|\mathbf{o}_{1:T}, \mathbf{a}_{1:T})}\right] d\mathbf{z}_{1:T}\, d\mathbf{y}_{1:T} \geq 0 \end{aligned} \tag{7}$$

applying Bayes rule to the term $p(\mathbf{z}_{1:T}, \mathbf{y}_{1:T}|\mathbf{o}_{1:T}, \mathbf{a}_{1:T})$ results in the following objective function

$$\begin{aligned} \log p(\mathbf{o}_{1:T}|\mathbf{a}_{1:T}) \geq\ & \mathbb{E}_{q_\theta(.)}[\log p(\mathbf{o}_{1:T}|\mathbf{z}_{1:T}, \mathbf{y}_{1:T}, \mathbf{a}_{1:T})] \\ & -\mathbb{E}_{q_\theta(.)}\left[\log\left(\frac{q_\theta(\mathbf{z}_{1:T}, \mathbf{y}_{1:T}|\mathbf{o}_{1:T}, \mathbf{a}_{1:T})}{p(\mathbf{z}_{1:T}, \mathbf{y}_{1:T}|\mathbf{a}_{1:T})}\right)\right] \end{aligned} \tag{8}$$

re-introducing Markov assumption into the regularization term of the ELBO

$$\begin{aligned} & q_\theta(\mathbf{z}_{1:T}, \mathbf{y}_{1:T}|\mathbf{o}_{1:T}, \mathbf{a}_{1:T}) = \\ & q_\theta(\mathbf{z}_1|\mathbf{y}_1) q_\phi(\mathbf{y}_1) \prod_{t=2}^{T} q_\theta(\mathbf{z}_t, \mathbf{y}_t|\mathbf{z}_{t-1}, \mathbf{y}_{t-1}, \mathbf{o}_{1:t}, \mathbf{a}_{1:t}) \end{aligned} \tag{9}$$

$$\begin{aligned} & q_\theta(\mathbf{z}_t, \mathbf{y}_t|\mathbf{z}_{t-1}, \mathbf{y}_{t-1}, \mathbf{o}_{1:t}, \mathbf{a}_{1:t}) = \\ & \frac{p(\mathbf{o}_t|\mathbf{z}_t) q_\theta(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{y}_t, \mathbf{a}_t) q_\phi(\mathbf{y}_t|\mathbf{z}_{t-1}, \mathbf{y}_{t-1}, \mathbf{a}_{t-1})}{\iint p(\mathbf{o}_t|\mathbf{z}_t) p(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{y}_t, \mathbf{a}_t) d\mathbf{z}_{t-1} d\mathbf{y}_t} \sim \\ & \eta \underbrace{q_\psi(\mathbf{z}_t|\mathbf{o}_t) q_\theta(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{y}_t, \mathbf{a}_t)}_{q^{filt}(\mathbf{z}_t)} q_\phi(\mathbf{y}_t|\mathbf{z}_{t-1}, \mathbf{y}_{t-1}, \mathbf{a}_{t-1}) \end{aligned} \tag{10}$$

where the denominator η is the normalization factor. The final ELBO:

$$\begin{aligned} \mathcal{F}_{\text{ELBO}}(\theta, \phi, \psi) = & \mathbb{E}_{q_\phi(.)}\left[\sum_{t=1}^{T} \log p(\mathbf{o}_t|\mathbf{z}_t)\right] \\ & - \beta(\sum_{t=2}^{T} \text{KL}[q^{filt}(\mathbf{z}_t)||p(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{y}_t, \mathbf{a}_t) \\ & - \sum_{t=2}^{T} \text{KL}[q_\phi(\mathbf{y}_t|.)||p(\mathbf{y}_t|\mathbf{a}_t, \mathbb{N})]) \end{aligned} \tag{11}$$

We employ the inverse variational measurement model $q_\psi(\mathbf{z}_t|\mathbf{o}_t)$ and perform Bayesian integration with the transition model $q_\theta(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{y}_t, \mathbf{a}_t)$ approximated by fully-connected network (FCN), to compute the filtered variational distribution $q^{filt}(\mathbf{z}_t)$. To approximate the indirectly observable distribution $q_\phi(\mathbf{y}_t|\mathbf{z}_{t-1}, \mathbf{y}_{t-1}, \mathbf{a}_{t-1})$, we utilize an LSTM network. Figure 2 presents illustration of the architecture and components of the filter. We adopt a β-regularization scheme, where β serves as a temperature-like parameter controlling the trade-off between reconstruction accuracy and regularization strength. To further improve generalization, we introduce a learnable

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

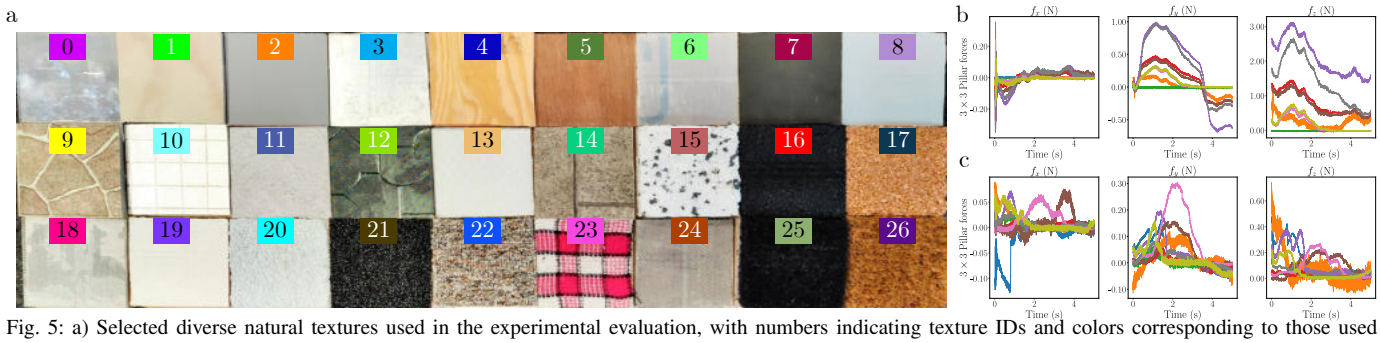

Fig. 5: a) Selected diverse natural textures used in the experimental evaluation, with numbers indicating texture IDs and colors corresponding to those used in the feature analysis figure. Illustration of raw tactile observations for Texture ID 0 (in b) and Texture ID 26 (in c) under same action parameters during the *vibrosliding* (vibration+sliding) interaction, demonstrating the distinct and information-rich tactile signals.

hierarchical prior $p(\mathbf{y}_t \mid \mathbf{a}_t, \mathbb{N})$, where $\mathbb{N}$ denotes the texture/object label. This prior acts as a structured regularizer: it encourages the latent variable $\mathbf{y}_t$ to encode causal physical properties associated with object identity and interaction context, while reducing the dependency on excessive tuning of β parameter. In the following section, we present the results of the proposed *Latent Filter*, highlighting its effectiveness in texture representation learning and analyzing the impact of vibration induced by the *Tacser* device during sliding interactions.

IV. Experiments

A. Robotic Setup, Data Collection & Training of Latent Filter

We evaluated our proposed approach for robotic texture perception on a dataset comprising 27 surface textures as shown in Figure 5. As everyday objects often lack sufficient variability, we curated a diverse texture set—drawing inspiration from [59], including samples from fabric, wood, plastic, and metal materials, exhibiting varying textural properties. The experiments were conducted with a robotic system equipped with the tactile sensor and vibrating mechanism (*Tacser*, Figure 3) mounted on a 6-DOF UR5 robotic arm (Figure 1). The system performed lateral sliding with vibration, referred to as *vibrosliding*, for 30 s. Since texture samples measured $10\times10\,\text{cm}^2$, the sliding direction was reversed at the edges, resulting in a zigzag motion along the x-axis, while vibrations were applied along the orthogonal y-axis. Initial contact was established using proportional velocity control along the z-axis, adjusting velocity until the mean normal force matched the desired contact force $f_{z_t} \in a_t$. In addition to the *vibroslide* interaction, we conducted static vibration (*vibration*) by activating the tacser without sliding ($v_{x_t} = 0$), and only lateral sliding (*sliding*) under the same contact force and velocity parameters but with vibration disabled ($d_{y_t} = 0$), to investigate the effectiveness of vibrosliding interaction. Each texture was probed in multiple regions to capture intra-sample variations, with 27 trials per texture under varying interaction parameters (three sliding velocities, three normal contact forces, and three repetitions), resulting in 729 interaction trajectories, each comprising 30,000 tactile observation frames. To address challenges in training with such long sequences, we applied two key transformations: we maintained the frequency information by using a resampling frequency of 800 Hz, and computed the $\log-mel$ spectrogram [60] using a window size of 800, a hop length of 801 and 128 mel bins. We then extracted the first 30 components of the mel bins, creating a spectrogram with dimensions of $30\times30\times30$. For the spatial component, we median-filtered and down-sampled the data to 30 Hz, combining the normal and shear data of each pillar of the tactile sensor and the global component to produce a dimension of $30\times30\times30$. Finally, the spatial and spectral matrices were concatenated to form the processed tactile observation data $\mathbf{o}_t$ (with n_o= $30\times30\times2$) (see inset Figure 2), and $T = 30$. For each texture, two of the three repeated trajectories were randomly assigned to train set and one to test set, for consistent evaluation across all baselines. The *Latent Filter* model was implemented in PyTorch and trained with the Adam optimizer (learning rate 10^{-5}), batch size 128, annealing β from $5{\cdot}10^{-3}$ to 10^{-1}, for 1000 epochs using the objective in Eq. 11.

B. Analysis of latent features

We evaluated the learned latent features (indirectly observable $\mathbf{y}_t$, dimension $n_y = 16$) against four baselines. Baseline I, adapted from Fisher et al. [25], approximated mechanical properties using statistical descriptors of time-series tactile data. Baseline II, based on Kaboli et al. [27], extracted 26-dimensional statistical features using Hjorth parameters. Baseline III employed a VAE model adapted from Aoyama et al. [61]. Baseline IV was an ablation of the proposed *Latent Filter* without the learnable hierarchical prior; this variant is closer in spirit to smoothing-based state-space models [52], though still structurally distinct due to the Bayesian integration component. For completeness, we also evaluated a standard sequential VAE (sVAE), which performed poorly, reinforcing why prior work on object property estimation has often relied on static VAEs. We conducted both qualitative and quantitative analyses. Figure 6 shows 3D UMAP embeddings [62] of features extracted by Baselines II–IV and our proposed method (final time-step), with 3D statistical features shown for Baseline I, across the three interaction settings. The figure highlights that unsupervised generative models achieve superior disentanglement and feature separation compared to handcrafted features (Baselines I and II). Furthermore, sequential models with action conditioning provide clearer clustering, with the proposed approach forming the most cohesive clusters. As UMAP does not preserve relative feature distances, we also computed the Euclidean distance of features to their object labels to assess inter- and intra-class separation (Figure 7). Distances were normalized by the maximum value

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

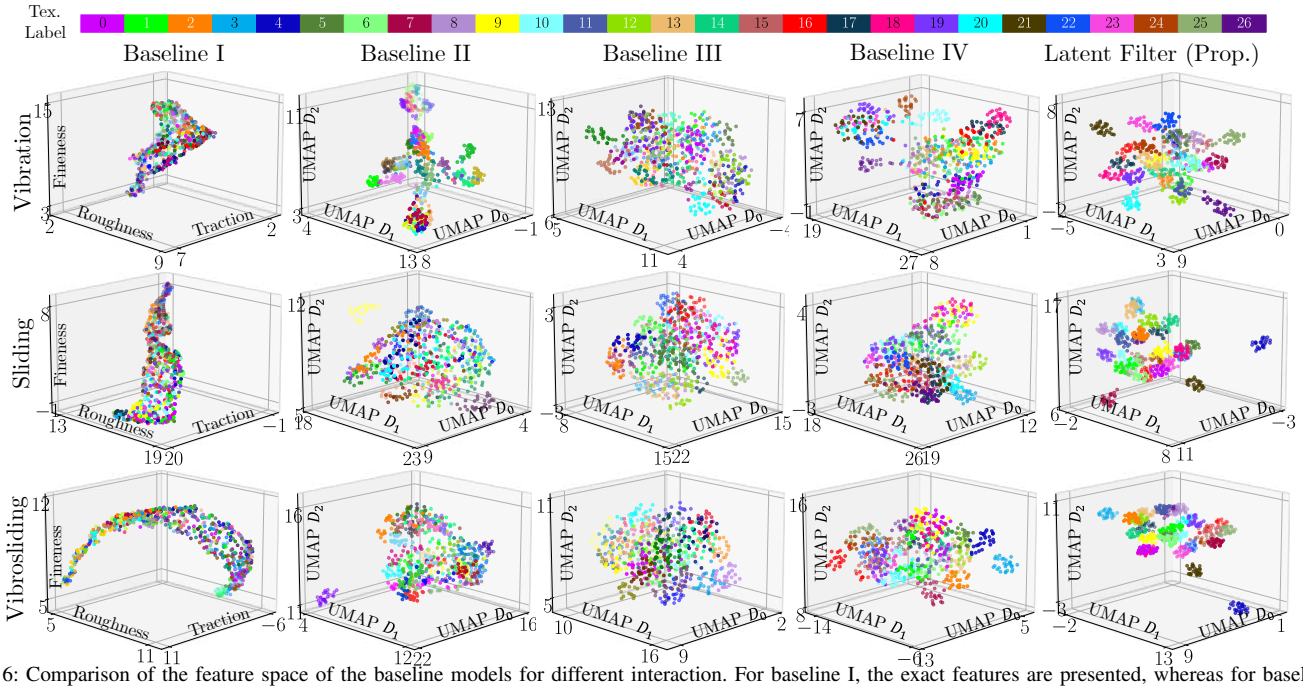

Fig. 6: Comparison of the feature space of the baseline models for different interaction. For baseline I, the exact features are presented, whereas for baseline II-IV and for our approach *latent filter* (last-time step) UMAP is utilized to project in 3D space. This provides insight into the discriminative capability of the features.

for comparability. Results show that our approach yields better class separation, with consistently low intra-class distances (blue). Notably, the proposed vibroslide interaction with tacser provided the strongest separation. For quantitative evaluation (Table I), we measured classification performance using logistic regression. To benchmark against discriminative modeling approaches, we also included a Frame-CNN model [31]. In addition, we computed the Silhouette L2 score [63] to assess clustering quality. Results clearly demonstrate that the proposed approach extracts meaningful mechanical properties, enabling reliable discrimination even with a linear classifier. Moreover, the vibrosliding interaction improves performance across all baselines, confirming that combining sliding with vibration enhances the richness of tactile observations.

TABLE I: Classification & Clustering Results (Higher value indicates better performance).

	Interaction	Base. I	Base. II	Base. III	Base. IV	FCNN	LF (Prop.)
Class.	*vibration*	4.10%	50.98%	69.57%	65.01%	78.00%	90.53%
	sliding	7.42%	47.53%	76.47%	76.45%	97.90%	99.00%
	vibrosliding	5.37%	57.86%	77.65%	84.7%	98.30%	100.00%
Cluster.	*vibration*	-0.3	-0.071	0.102	0.083	NA	0.41
	sliding	-0.275	-0.171	0.126	0.132	NA	0.508
	vibroslide	-0.255	-0.152	0.114	0.208	NA	0.534

C. Analysis of sequential inference

To further evaluate the strength of sequential inference, we fitted a non-linear kernel ridge regression model [64] to map the latent features $\mathbf{y}_t$ to normalized texture labels. The regressor was trained on the final time-step features $\mathbf{y}_T$ and then applied to samples drawn from the variational distribution $q_\psi(\mathbf{y}_t|\cdot)$, $t = 0, \ldots, T-1$. Although such mappings could ideally be extended to estimate characterizable physical properties (e.g., Coulomb friction), this remains challenging for natural textures. Prediction accuracy was measured using NMSE between predicted and true labels. Figure 8 shows the temporal evolution of NMSE for Baseline IV and the proposed approach (the only sequential models) across interaction types (*vibration*, *sliding*, *vibrosliding*). We highlight three representative parameter settings (low, intermediate, and high values of sliding speed and normal force, where applicable). Results demonstrate that *vibrosliding* accelerates the convergence of latent features, improves accuracy and reduces sensitivity to interaction parameters, outperforming *vibration* and *sliding* interactions.

V. Discussion & Conclusion

This letter has introduced a novel approach *Latent Filter* to infer the properties of diverse textures and surfaces. Unsupervised deep state-space modeling and training overcome the limitations of class-specific texture recognition found in prior work. We demonstrate that incorporating sequential modeling and action conditioning significantly enhances feature extraction compared to static models such as baseline III. Furthermore, the hierarchical prior improves performance over the ablative Baseline IV. Importantly, this prior is applied only during training and does not constrain the generative model's generalization during inference, unlike the strict class conditioning used in discriminative models. In addition, we explored a novel texture exploration approach via the combination of vibration and sliding using a novel device *Tacser*, which facilitated the disentanglement of features and led to more robust representations, as shown in Figure 8, Figure 7, and Table. I. From the feature distance analysis, we observe that vibration interactions are particularly effective in discriminating rough textures, while sliding excels at capturing frictional properties. This highlights the strength of the proposed *Tacser* design, which enables orthogonal exploration and thereby leverages the complementary advantages of both interaction. Although the performance of vibration-only interactions was

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

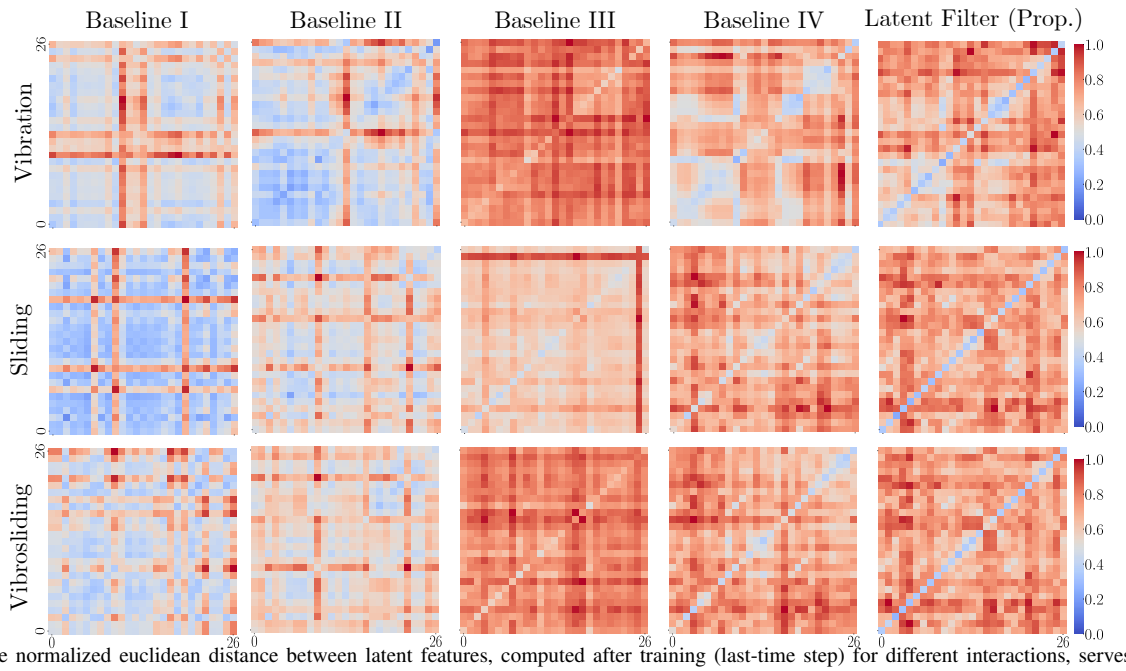

Fig. 7: The normalized euclidean distance between latent features, computed after training (last-time step) for different interactions, serves as a measure of similarity in the inferred textural properties. A smaller Euclidean distance (blue hue) indicates greater similarity in the underlying feature characteristics of the textures.

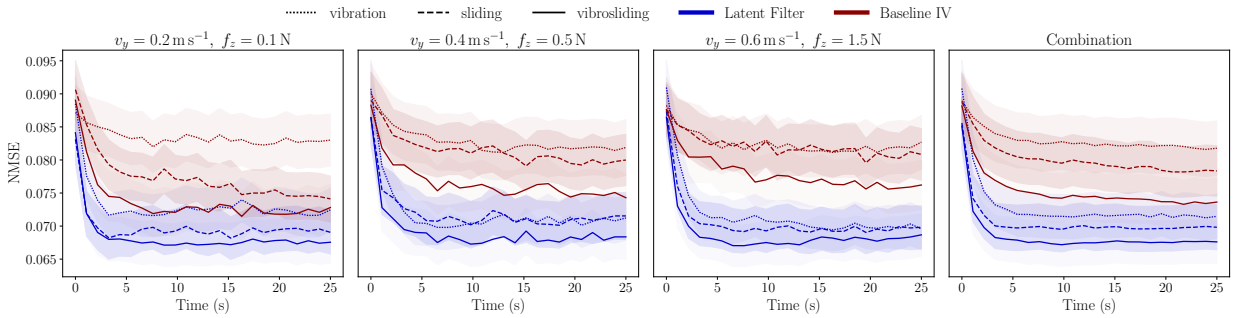

Fig. 8: Evolution of the NMSE values over time.

lower, the proposed Latent Filter still outperformed all baseline approaches, making it a viable option in scenarios where surfaces are non-planar or lateral sliding is impractical (e.g., grasping curved objects). Additionally, we illustrated how the learned features can be utilized effectively with lightweight classifiers such as Logistic Regression for texture discrimination (Table I). This highlights the model's capability to identify causal physical properties (invariants) instead of just categorizing predefined labels. This has profound implications for downstream control, where simple learned or analytical functions can directly exploit continuous, differentiable latent space for tasks such has slip avoidance of complex textural objects. In contrast, class-based discretization, even at fine resolution inevitable introduces discontinuities, and forces the use of inefficient lookup-based strategies. In the future, a carefully constructed artificial texture set will have to be created to further study the correlation between the distance in the feature space and the physical properties. Moreover, additional constraints could be introduced in the deep state-space modeling to ensure that the extracted features are better aligned with the structure of the Euclidean space [65]. The selected tactile sensor effectively captured rich spatio-temporal tactile features with high resolution and suitable mechanical properties for high-frequency vibrations. As a future direction, it will be valuable to compare with diverse tactile sensors, including vision-based ones, to examine whether higher spatial resolution and static interaction can reduce or replace the need for dynamic interaction. Beyond texture perception, our approach can be extended to more complex applications, such as soft-object interaction, where the interaction is complex to model/characterize. Moreover, this technique can serve as a general robust framework for model learning with potential applications in planning and control. In particular, the uncertainty estimates and relative changes in latent values provided by our method can be used for fine-grained, closed-loop control such as slip-avoidance. In conclusion, this study significantly improves robotic texture perception by highlighting the role of interaction beyond tactile sensing technology and action-conditioned modeling.

REFERENCES

[1] A. Dutta *et al.*, "Predictive visuo-tactile interactive perception framework for object properties inference," *IEEE Trans. Robo.*, 2025.
[2] R. Bajcsy *et al.*, "Revisiting active perception," *Auto. Robi,*, 2018.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

[3] A. Dutta *et al.*, "Vitract: robust object shape perception via active visuo-tactile interaction," *IEEE RA-L*, 2024.

[4] J. Bohg *et al.*, "Interactive perception: Leveraging action in perception and perception in action," *IEEE Trans. Rob.*, 2017.

[5] X. Song *et al.*, "Efficient break-away friction ratio and slip prediction based on haptic surface exploration," *IEEE Trans. Robo.*, 2013.

[6] B. Sundaralingam and T. Hermans, "In-hand object-dynamics inference using tactile fingertips," *IEEE Trans. Robo.*, 2021.

[7] Q. Li *et al.*, "A review of tactile information: Perception and action through touch," *IEEE Trans. Robot.*, 2020.

[8] L. Yu and D. Liu, "Recent progress in tactile sensing and machine learning for texture perception in humanoid robotics," *Inter. Mat.*, 2025.

[9] G. A. Waltersson and Y. Karayiannidis, "Planar friction modelling with lugre dynamics and limit surfaces," *IEEE Trans. Robo.*, 2024.

[10] V. Hayward, "Is there a 'plenhaptic'function?" *Phil.l Trans. Royal Society B: Biological Sciences*, 2011.

[11] S. J. Lederman and R. L. Klatzky, "Hand movements: A window into haptic object recognition," *Cog. Psycho.*, 1987.

[12] S. J. Lederman and R. L. Klatzky, "Extracting object properties through haptic exploration," *Acta Psycho.*, 1993.

[13] V. Chu *et al.*, "Using robotic exploratory procedures to learn the meaning of haptic adjectives," in *Proc. Inte. Conf. Robo. and Auto.*, 2013.

[14] D. Katz and L. E. Krueger, *The world of touch*. Psychology press, 2013.

[15] M. Hollins and S. R. Risner, "Evidence for the duplex theory of tactile texture perception," *Perception & psychophysics*, vol. 62, no. 4, pp. 695–705, 2000.

[16] U. B. Rongala *et al.*, "The import of skin tissue dynamics in tactile sensing," *Cell Reports Physical Science*, vol. 5, no. 5, 2024.

[17] Y. Kurita *et al.*, "Wearable sensorimotor enhancer for a fingertip based on stochastic resonance," in *Proc. Inte. Conf. Robo. and Auto.* IEEE, 2011.

[18] Y. Tada, K. Hosoda, and M. Asada, "Sensing ability of anthropomorphic fingertip with multi-modal sensors," in *Proc. Inte. Conf. Intell. Robot. and Sys.* IEEE, 2004.

[19] T. Araki *et al.*, "Experimental investigation of surface identification ability of a low-profile fabric tactile sensor," in *Proc. Inte. Conf. Intell. Robot. and Sys.* IEEE, 2012.

[20] N. Jamali *et al.*, "Texture recognition by tactile sensing," in *Proc. Austral. Conf. Robo. and Auto. (ACRA)*, 2009.

[21] C. M. Oddo *et al.*, "Roughness encoding for discrimination of surfaces in artificial active-touch," *IEEE Trans. Robo.*, 2011.

[22] N. Bai *et al.*, "A robotic sensory system with high spatiotemporal resolution for texture recognition," *Nature Comm.*, vol. 14, no. 1, p. 7121, 2023.

[23] P. Giguere and G. Dudek, "A simple tactile probe for surface identification by mobile robots," *IEEE Trans. Robo.*, 2011.

[24] J. Hoelscher *et al.*, "Evaluation of tactile feature extraction for interactive object recognition," in *Proc. Inte. Conf. on Humanoid Robo. (Humanoids)*. IEEE, 2015.

[25] J. A. Fishel and G. E. Loeb, "Bayesian exploration for intelligent identification of textures," *Front. in Neuro.*, 2012.

[26] V. a. Chu, "Robotic learning of haptic adjectives through physical interaction," *Robo. and Auto. Systems*, 2015.

[27] M. Kaboli and G. Cheng, "Robust tactile descriptors for discriminating objects from textural properties via artificial robotic skin," *IEEE Trans. on Rob.*, 2018.

[28] P. Uttayopas *et al.*, "Object recognition using mechanical impact, viscoelasticity, and surface friction during interaction," *IEEE Trans. on Haptics*, vol. 16, no. 2, pp. 251–260, 2023.

[29] L. Liu *et al.*, "From bow to cnn: Two decades of texture representation for texture classification," *Inte. Jour. Comp. Vision*, 2019.

[30] M. Meribout *et al.*, "Tactile sensors: A review," *Measurement*, 2024.

[31] Y. Massalim *et al.*, "Deep vibro-tactile perception for simultaneous texture identification, slip detection, and speed estimation," *Sensors*, 2020.

[32] T. Taunyazov *et al.*, "Towards effective tactile identification of textures using a hybrid touch approach," in *Proc. Inte. Conf. Robo. and Auto.* IEEE, 2019.

[33] S. S. Baishya and B. Bäuml, "Robust material classification with a tactile skin using deep learning," in *Proc. Inte. Conf. Intell. Robot. and Sys.* IEEE, 2016.

[34] J. a. Wei, "Multimodal unknown surface material classification and its application to physical reasoning," *IEEE Trans. Industrial Inform¿*, 2021.

[35] Y. a. Gao, "Deep learning for tactile understanding from visual and haptic data," in *Proc. Inte. Conf. Robo. and Auto.* IEEE, 2016.

[36] N. Pestell and N. F. Lepora, "Artificial sa-i, ra-i and ra-ii/vibrotactile afferents for tactile sensing of texture," *Jour. Royal Society Interface*, 2022.

[37] G. Cao *et al.*, "Spatio-temporal attention model for tactile texture recognition," in *Proc. Inte. Conf. Intell. Robot. and Sys.* IEEE, 2020.

[38] W. Yuan *et al.*, "Connecting look and feel: Associating the visual and tactile properties of physical materials," in *Proc. IEEE Conf. Comp. Vis. and Pattern Recog.*, 2017.

[39] V. Dave *et al.*, "Multimodal visual-tactile representation learning through self-supervised contrastive pre-training," in *Proc. Inte. Conf. Robo. and Auto.*, 2024.

[40] Q. Xi *et al.*, "Cm-avae: Cross-modal adversarial variational autoencoder for visual-to-tactile data generation," *IEEE Robo. and Auto. Lett.*, 2024.

[41] M. Purri and K. Dana, "Teaching cameras to feel: Estimating tactile physical properties of surfaces from images," in *ECCV*. Springer, 2020.

[42] Y. Fang *et al.*, "Bidirectional visual-tactile cross-modal generation using latent feature space flow model," *Neural Networks*, 2024.

[43] Q. Yang *et al.*, "Drop to transfer: Learning transferable features for robot tactile material recognition in open scene," *IEEE Trans. Inst. and Meas.*, 2023.

[44] K. Liu, Q. Yang, Y. Xie, and X. Huang, "Towards open-set material recognition using robot tactile sensing," in *Proc. Inte. Conf. Robo. and Auto.* IEEE, 2023.

[45] G. Cao *et al.*, "Multimodal zero-shot learning for tactile texture recognition," *Robo. and Auto. Sys.*, 2024.

[46] J. Edwards *et al.*, "Extracting textural features from tactile sensors," *Bioinspiration & biomimetics*, 2008.

[47] S.-a. Wang *et al.*, "Fabric classification using a finger-shaped tactile sensor via robotic sliding," *Front. in Neuro.*, 2022.

[48] Z. Shao *et al.*, "Haptic recognition of texture surfaces using semi-supervised feature learning based on sparse representation," *Cognitive Comput.*, 2023.

[49] C. Higuera *et al.*, "Sparsh: Self-supervised touch representations for vision-based tactile sensing," *arXiv preprint arXiv:2410.24090*, 2024.

[50] F. Yang *et al.*, "Touch and go: Learning from human-collected vision and touch," in *Adv. Neural Info. Proc. Sys.* Curran Associates, Inc., 2022.

[51] A. Klushyn *et al.*, "Latent matters: Learning deep state-space models," in *Adv. Neural Info. Proc. Sys.* Curran Associates, Inc., 2021.

[52] P. Becker-Ehmck, "Latent state-space models for control," Ph.D. dissertation, Technische Universität Darmstadt, Darmstadt, 2022.

[53] N. Heravi *et al.*, "Development and evaluation of a learning-based model for real-time haptic texture rendering," *IEEE Trans. Haptics*, 2024.

[54] N. Komeno and T. Matsubara, "Tactile perception based on injected vibration in soft sensor," *IEEE Robo. Auto. Lett.*, 2021.

[55] Contactile, "Contactile," https://contactile.com/, 2022, [Online].

[56] Actronika, "Hapcoil," https://uploads-ssl.webflow.com/5eb037130a8b570a78e002a0/600946155d92f4663545dbfa_HapCoil_One_HC121238O_datasheet.pdf.

[57] M. Holliins *et al.*, "Perceptual dimensions of tactile surface texture: A multidimensional scaling analysis," *Perception & psychophysics*, 1993.

[58] L. Ljung and T. Glad, *Modeling of dynamic systems*. USA: Prentice-Hall, Inc., 1994.

[59] M. Strese *et al.*, "A haptic texture database for tool-mediated texture recognition and classification," in *Int. Symp. on Haptic, Audio and Visual Envir. and Games Proce.* IEEE, 2014, pp. 118–123.

[60] B. McFee *et al.*, "librosa: Audio and music signal analysis in python." in *SciPy*, 2015, pp. 18–24.

[61] M. Y. Aoyama *et al.*, "Few-shot learning of force-based motions from demonstration through pre-training of haptic representation," in *Proc. Inte. Conf. Robo. and Auto.*, 2024.

[62] L. McInnes *et al.*, "Umap: Uniform manifold approximation and projection," *Journal of Open Source Software*, 2018.

[63] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, 1987.

[64] K. P. Murphy, "Kernel ridge regression," in *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: The MIT Press, 2012, ch. 14.4.3, pp. 492–493.

[65] N. Chen *et al.*, "Learning flat latent manifolds with VAEs," in *Proc. 37th Int. Conf. on Machine Learn.* PMLR, 2020, pp. 1587–1596.