IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 June 2025, accepted 16 June 2025, date of publication 20 June 2025, date of current version 2 July 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3581685

RESEARCH ARTICLE

Estimating Perceptual Attributes of Haptic Textures Using Visuo-Tactile Data

MUDASSIR IBRAHIM AWAN[1], (Graduate Student Member, IEEE), AND SEOKHEE JEON[2]
[1]Department of Computer Engineering, Kyung Hee Uinversity, Gyeonggi, Yongin 17104, South Korea
[2]Department of Metaverse, Kyung Hee Uinversity, Gyeonggi, Yongin 17104, South Korea

Corresponding author: Seokhee Jeon (jeon@khu.ac.kr)

This research was supported by the Ministry of Science and Information & Communications Technology (ICT), Korea, in part through the Institute for Information & communications Technology Planning & evaluation (IITP) under the Graduate School of Metaverse Convergence Support Program (IITP-2024-RS-2024-00425383), and in part through the National Research Foundation of Korea (NRF) under the Mid-Researcher Program (2022R1A2C1008483).

ABSTRACT Accurate prediction of perceptual attributes of haptic textures is essential for advancing VR and AR applications and enhancing robotic interaction with physical surfaces. This paper presents a deep learning-based multi-modal framework, incorporating visual and tactile data, to predict perceptual texture ratings by leveraging multi-feature inputs. To achieve this, a four-dimensional haptic perceptual space encompassing rough-smooth, flat-bumpy, sticky-slippery, and hard-soft dimensions is first constructed through psychophysical experiments, where participants evaluate 50 diverse real-world texture samples. A physical signal space is subsequently created by collecting visual and tactile data from these textures. Finally, a deep learning architecture integrating a CNN-based autoencoder for visual feature learning and a ConvLSTM network for tactile data processing is trained to predict user-assigned attribute ratings. This multi-modal, multi-feature approach maps physical signals to perceptual ratings, enabling accurate predictions for unseen textures. To evaluate predictive accuracy, we employed leave-one-out cross-validation to rigorously assess the model's reliability and generalizability against several machine learning and deep learning baselines. Experimental results demonstrate that the framework consistently outperforms single-modality approaches, achieving lower MAE and RMSE, highlighting the efficacy of combining visual and tactile modalities.

INDEX TERMS Haptic texture attributes, visuo-tactile learning, deep learning, tactile signal processing, texture recognition, human haptic perception.

I. INTRODUCTION

When a textured surface is stroked, a range of tactile signals is generated, playing a crucial role in surface texture perception. Humans interpret these signals rapidly within a multi-dimensional haptic perceptual space, characterized by descriptors such as roughness, softness, and bumpiness [1]. This cognitive process, driven by deformation of the user's skin connected to the user's interactions, enables material recognition and object identification with remarkable accuracy [2], [3]. Computationally modeling this perception is essential for applications in virtual reality (VR), augmented reality (AR), haptic-enabled robotics, and human-computer interaction [4].

Primary research on haptic textures has focused on modeling surface interactions and generating realistic haptic feedback. A number of studies have introduced haptic texture rendering modules and libraries that synthesize acceleration signals of specific textures based on physical interactions [5], [6]. These modules provide a means to generate haptic feedback for virtual environments. However, while the synthesis of texture signals has been well studied, the question of when to use these models effectively remains largely unexplored. Understanding their applicability is crucial for improving haptic rendering fidelity, particularly in scenarios requiring accurate perceptual predictions.

One of the key applications where accurate haptic attribute prediction is essential is model-mediated teleoperation, where a remote system captures interaction data, but real-time transmission of raw haptic signals may be limited due

The associate editor coordinating the review of this manuscript and approving it for publication was Brian Ng.

to latency. In such cases, a perceptually aligned haptic texture model can be selected from a texture library [5], ensuring that the reconstructed feedback on the operator's side closely matches the intended material properties, thereby enhancing realism in remote interactions [7]. These systems utilize a rigid tool equipped with vision and tactile sensors, where the vision sensor captures global texture characteristics, such as macro-scale structure and material reflectance, while the tactile sensor records tool-induced vibrations as high-frequency acceleration signals, along with scanning speed and applied force. However, to ensure that the remotely rendered haptic feedback is perceptually consistent with the original material, accurate haptic attribute prediction is needed to establish a reliable mapping between physical signals and human perception. By leveraging predictive models, the system can generate an appropriate haptic representation rather than relying on direct signal transmission that may be affected by latency or interaction variability [7], [8].

To better conceptualize the idea, two fundamental representational spaces are introduced: the perceptual attribute space and the physical signal space. The perceptual, or haptic, attribute space is constructed through psychophysical experiments, where participants rate textures along bipolar attributes such as rough–smooth and hard–soft, forming a subjective representation of haptic perception [9], [10].

In contrast, the physical signal space is derived from measured texture characteristics, including high-resolution visual data and tactile signals such as acceleration, applied force, and scanning speed, which collectively encode the objective physical properties of textures [11], [12]. Establishing a reliable mapping between these spaces is essential for computational models to predict how a given texture will be perceived based on its physical attributes.

While humans use both the visual and tactile data to gauge the haptic texture, most previous studies on the prediction of haptic texture perceptual attributes rely on single-modality approaches, using either tactile signals or visual texture analysis. On one hand, while tactile data provide high accuracy in capturing micro-textural properties, it is highly sensitive to interaction parameters, such as force, speed, and sensor noise, which can introduce inconsistencies in estimation. On the other hand, visual data possess macro-scale structural patterns but do not provide compliance-related attributes, such as softness or friction, which are not always visually discernible [2], [7]. Visual data also do not have micro-scale information that greatly influences texture perception, due to the spatial resolution of the visual sensors. Some studies have attempted to integrate multi-modal learning, but they predominantly focus on texture classification rather than continuous haptic attribute prediction, limiting their ability to model perceptual variations accurately [7], [11], [13]. Since vision and tactile data encode complementary information, a robust multi-modal framework that effectively fuses both modalities is necessary to improve haptic attribute prediction, enhance generalizability, and ensure perceptual alignment with human ratings.

Nonetheless, while various computational techniques have been explored to map perceptual attributes from physical signals, challenges remain in achieving robust and generalizable haptic prediction. Early attempts employed parametric models but often struggle to generalize across different textures and interaction conditions [2]. More recent efforts have leveraged deep learning-based models to learn complex mappings between input signals and haptic attributes [14], [15]. These methods have demonstrated notable success, particularly in capturing intricate texture representations and improving prediction accuracy. However, many existing deep learning models are still trained on single-modality data, which can limit their ability to generalize across diverse textures. Multi-modal deep learning approaches that integrate visual and tactile data have shown promising results, with different fusion strategies being explored to enhance their effectiveness. One promising direction involves cross-modal feature fusion, where representations extracted from different sensory modalities are effectively combined to improve prediction accuracy [7], [11]. Additionally, leveraging features extracted from pre-trained models and integrating them with classical handcrafted descriptors provides a robust way to capture both high-level abstract features and fine-grained physical properties, further enhancing the reliability of haptic attribute prediction [15].

Motivated by these challenges, this work introduces a deep learning framework that integrates visual and tactile data for predicting perceptual haptic attributes. The framework constructs a physical signal space by capturing high-resolution images and tactile data, including acceleration, applied force, and scanning speed from 50 real-world textures, while simultaneously establishing a four-dimensional perceptual space through psychophysical experiments, where participants rate textures along the bipolar attributes of rough-smooth, flat-bumpy, sticky-slippery, and hard-soft. To facilitate the analysis of these perceptual ratings, the four-dimensional space is visualized in a two-dimensional representation, providing insights into the structure of human haptic perception. To establish a mapping between these spaces, the framework employs a two-stream architecture, where the visual stream extracts texture features using a CNN-based autoencoder, incorporating pre-trained ResNet-50 [16] features alongside Gray-Level Co-occurrence Matrix (GLCM) descriptors to enhance structural representation. The tactile stream, implemented as a Convolutional LSTM (ConvLSTM) network [17], processes high-frequency vibration signals using Mel-Frequency Cepstral Coefficients (MFCCs), complemented by interaction parameters such as scanning speed and applied force to improve robustness. By integrating these complementary modalities, the framework strengthens feature representation and enhances perceptual alignment, leading to more accurate haptic attribute predictions.

Beyond applications in teleoperation, haptic attribute prediction can also serve as a scalable alternative to human perceptual evaluation. Directly assessing texture properties through psychophysical experiments is often impractical due to time, cost, and logistical constraints, while in certain cases, such as analyzing hazardous surfaces or conducting large-scale material perception studies, direct human interaction is infeasible [9], [10]. By leveraging multimodal sensory data, a predictive model can enable efficient, reproducible, and scalable estimation of perceptual haptic attributes, reducing the dependency on resource-intensive human studies while maintaining perceptual alignment. Another key application is perception-based data compression and transmission. Instead of storing and transmitting raw physical data, perceptual attributes can be estimated from newly captured signals, encoded with compression, and efficiently stored or transmitted. This approach can significantly reduce data storage requirements and transmission bandwidth. Eventually, rendering algorithms and haptic devices can convert perceptual attribute values into appropriate commands or physical signals tailored to the user's interaction.

The primary contributions of this work are as follows:

- Development of a dual-stream multi-modal deep learning framework for haptic attribute prediction, combining a CNN-based autoencoder for visual feature encoding with a ConvLSTM network for modeling temporal tactile signals.
- Collection of a multi-modal texture dataset consisting of high-resolution images and tactile signals from 50 unique texture surfaces, including acceleration, applied force, and scanning speed.
- Structuring and visualizing a four-dimensional perceptual space using bipolar haptic attributes (rough–smooth, flat–bumpy, sticky–slippery, and hard–soft) to improve the interpretability and spatial organization of textures based on subjective ratings.
- Quantitative evaluation using LOOCV and comparison against baseline models, demonstrating improved prediction accuracy for each perceptual attribute.

The paper is organized as follows. Section II provides a review of related work. The proposed method, including the architecture of the attribute prediction model and its input-output schema, is introduced in Section III. Section IV describes the construction of the haptic perceptual space. The collection and preprocessing of visual-tactile data, which form the basis of the physical feature space, are outlined in Section V. Evaluation procedures and results are presented in Section VI, followed by a discussion of the framework in Section VII. Finally, Section VIII concludes the study.

II. RELATED WORKS

Below, we discuss related work on haptic texture attributes and their organization within perceptual spaces, the use of tactile and visual data, as well as deep learning approaches for texture analysis.

A. HAPTIC TEXTURE ATTRIBUTES AND PERCEPTUAL SPACES

Haptic texture attributes are perceptual qualities humans associate with surfaces, such as roughness, and slipperiness. These attributes can be perceived through bare-finger or tool-based interaction and form the basis for constructing haptic perceptual spaces, multidimensional representations that characterize textures by human perception [18].

One of the pioneering studies on understanding haptic texture perception was conducted by Yoshida et al. [19], focusing on bare-finger interactions. Their research identified four key perceptual dimensions of texture: hard-soft, heavy-light, cold-warm, and rough-smooth. Subsequent work refined these findings, confirming rough–smooth and hard–soft as dominant bipolar dimensions [20]. Further expansion introduced macro- and micro-roughness as distinct perceptual axes and highlighted friction as a critical attribute [21]. Collectively, these studies contributed to the widely recognized five perceptual dimensions: micro-roughness, macro-roughness, friction, stiffness, and warmth. In contrast, tool-mediated methods, such as those employed by [22], demonstrated that tapping with a rigid probe enhances the perception of hardness and softness. Other studies demonstrated that tool-based interactions reliably assess the rough–smooth dimension by reducing variability in skin contact [18], [23]. Despite their effectiveness, tool-mediated approaches may fail to capture finer details like friction and micro-roughness. In these cases, bare-finger interactions provide richer and more nuanced feedback, which is essential for accurately capturing subtle surface properties [24].

Haptic attributes derived from user interactions are commonly used to construct perceptual spaces that characterize the multi-dimensional nature of texture perception. These spaces are typically generated using techniques like Multi-Dimensional Scaling (MDS) [18] or Principal Component Analysis (PCA) [25], [26], which reduce dimensionality for easier interpretation. Perceptual spaces play a crucial role in texture analysis [6], [27] and the development of virtual textures [28]. While dimensionality reduction simplifies data, it can overlook important perceptual details and is unsuitable for estimating actual human-assigned ratings. In our recent work [15], we introduced a four-dimensional haptic perceptual space consisting of two 2D subspaces, preserving raw user ratings without reducing dimensions. This approach offers a more accurate and detailed representation of perceptual attributes, directly reflecting the degree of user-assigned ratings for each texture dimension. Despite progress, further research is needed to develop intuitive representations that incorporate actual user ratings, enhancing the understanding of haptic perception.

B. TACTILE AND VISION DATA FOR TEXTURE ANALYSIS

Texture analysis through tactile feedback involves capturing the unique vibrations generated when interacting with surfaces. This feedback reflects both macro features (e.g., bumpiness) and micro features (e.g., fine roughness) [22], [24]. However, the relationship between texture properties, user motion, and the resulting vibrations is inherently complex and nonlinear, posing significant challenges for accurate modeling and distinguishing [29]. Early studies recorded tactile data under fixed interaction parameters or segmented it into stationary signals, limiting generalizability [7]. Recent approaches have shifted towards directly utilizing data collected through free-hand motion, preserving natural interactions and a wider range of vibratory responses [30] without information loss. However, even with parametric and deep learning models, distinguishing similar textures remains difficult due to overlapping vibratory signals [2], [14], [30].

In contrast,Vision-based techniques offer a simpler alternative to tactile sensors, requiring less specialized hardware. Heravi et al. [31] used GelSight images to classify textures effectively, though their focus was on texture rendering rather than precise attribute prediction. Yang et al. [32] aligned GelSight embeddings with visual and auditory modalities to improve classification accuracy. Similarly, Luo et al. [33] employed the ViTac dataset, which combines camera and GelSight data, to classify cloth textures. The MIT GelSight dataset [34] has also been widely adopted for high-resolution surface geometry analysis and fine-grained texture recognition. Although these methods and datasets support valuable classification tasks, they do not provide user-rated perceptual labels, limiting their applicability for modeling subjective haptic impressions. While, Hassan et al. [15] employed a feature-based approach using texture images to estimate haptic attributes, showing strong results but struggling in predicting attributes like softness and fine roughness. This is likely because image-based methods primarily capture macro features (e.g., surface patterns) but often miss sub-surface features (e.g., material compliance / softness), limiting their accuracy in similar applications.

These challenges are well-recognized in the haptics community. To address them, studies have explored integrating visual and tactile features for more robust texture analysis [25]. Fusing visual data, which captures macro features, with tactile data, which reflects micro details, enables a comprehensive representation of textures. This multi-modal approach improves the prediction of perceptual attributes, surpassing simple texture classification [12], [35]. By leveraging shared features from both modalities, models achieve better generalization and accuracy, even for unseen textures. However, most efforts focus on classification, with limited research addressing regression for predicting perceptual haptic attributes [12], [36].

C. DEEP LEARNING APPROACHES FOR TEXTURE PERCEPTION

Recent studies have increasingly adopted deep learning (DL) approaches for modeling texture perception from both visual and tactile data [14], [35]. Convolutional Neural Networks (CNNs) are commonly used for extracting spatial features from visual textures, effectively capturing structural and geometric patterns in images, and have shown significant performance in texture recognition tasks [15]. Tactile signals, in contrast, are inherently spatio-temporal, as they contain both local surface-related information and dynamic variations over time. In earlier works, researchers applied CNN-based models to time-series tactile data, focusing on extracting local features from vibration signals [12]. With the introduction of Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks that are specialized in modeling temporal dependencies, researchers began to explore CNN-LSTM hybrid architectures. These models typically combine CNN and LSTM in a stacked or two-stream configuration, where CNNs are used to extract spatial features from input segments and LSTMs capture temporal dynamics across those segments [14]. While this setup enables joint spatio-temporal learning, it often introduces challenges including complex optimization, sensitivity to hyperparameters, and reduced spatial coherence when features are temporally sequenced. These limitations are not exclusive to haptic data and have been widely observed in other spatio-temporal learning tasks [14], [37], [38].

To address the limitations of existing architectures for processing time-series data, recent works have explored several alternatives, with Transformer frameworks [39] and Convolutional LSTM (ConvLSTM) networks [17] being among the most prominent. Transformers perform well in sequential tasks but are constrained by large data needs and can limit their practicality in texture-based haptic tasks [40]. ConvLSTM, however, offers a structured approach for modeling both spatial and temporal dependencies and has demonstrated effectiveness in a variety of time-series domains. Notably, Zhang et al. [41] applied ConvLSTM to tactile sensing through the FingerVision system, enabling slip detection and object recognition by capturing spatiotemporal dynamics in tactile signals. These results support the suitability of ConvLSTM for haptic perception tasks involving dense sensor input with temporal variation.

Despite its success in related domains, ConvLSTM remains unexplored for haptic texture analysis. We hypothesize that it is better suited for learning the underlying spatial and temporal structure of tactile signals compared to LSTMs or CNN-LSTM hybrids.For the visual modality in our multimodal framework, we consider CNNs an effective choice for extracting spatial features, including local texture patterns and geometric structures, given their proven ability to capture these characteristics in texture images.

III. VISUO-TACTILE NET

The primary objective of this work is to predict haptic affective attributes from multimodal physical signals using a structured computational framework. As shown in Figure 1, the process begins with the preparation of texture samples obtained from real-world surfaces. Next, we construct two distinct data spaces: 1) Physical Feature Space (PFS), which includes multimodal physical signals captured from the texture samples, and 2) Haptic Perceptual Space (HPS), which contains user-assigned perceptual attribute ratings

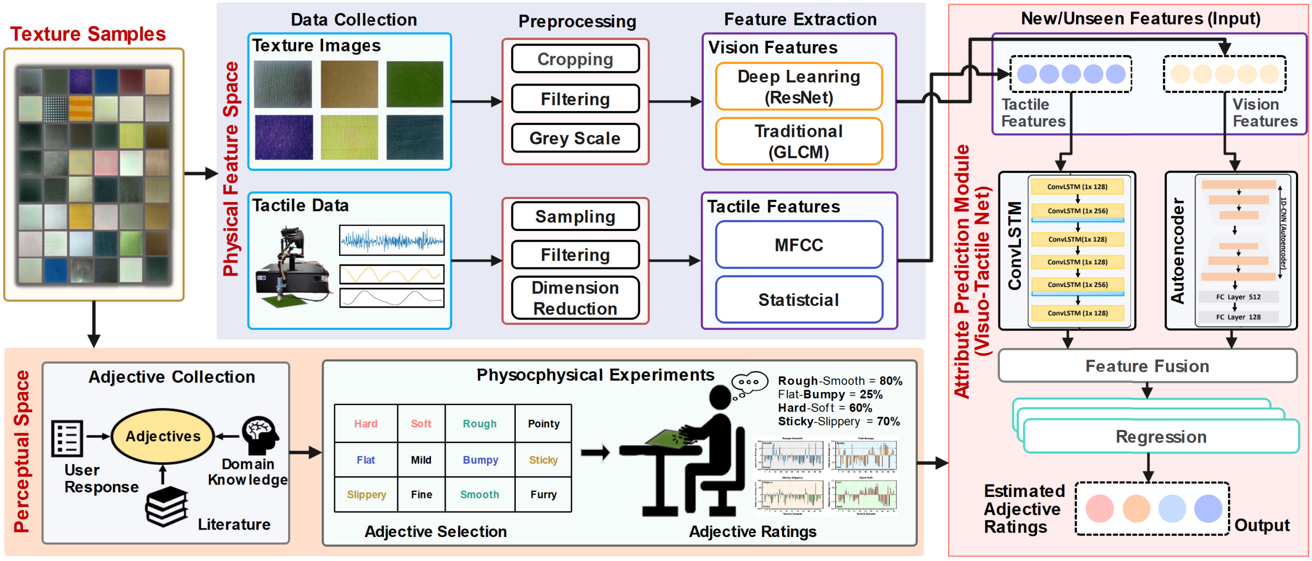

FIGURE 1. Overall framework.

gathered through psychophysical experiments. The core of this framework is the Visuo-Tactile Net, which bridges the gap between physical signals and human perceptual ratings. Its design is detailed in the remainder of this section, while the construction of the HPS and PFS is described in Sections IV and V, respectively.

To facilitate the mapping between physical features and perceptual attributes, the Visuo-Tactile Net employs a dual-stream architecture which we also termed as Visuo-tactile Net (Figure 2) consisting of two parallel branches: the Haptic Vision Network (HV-Net) and the Haptic Tactile Network (HT-Net). Both networks operate on pre-extracted features rather than raw data to improve robustness and mitigate overfitting (see Sections V and IV). HV-Net encodes visual information from texture images, while HT-Net models temporal patterns in tactile signals. Finally, the dual-stream architecture fuses visual and tactile features from HV-Net and HT-Net to create a robust joint representation of physical texture, which is then used to predict haptic attributes. The following subsections describe the design of each stream and the associated training methodology.

A. HAPTIC VISION NETWORK (HV-NET)

The HV-Net generates compact and discriminative representations from visual texture inputs by integrating deep and statistical features. The input to HV-Net combines high-level descriptors extracted using a pretrained ResNet-50 model [16] with handcrafted texture descriptors derived from the Gray-Level Co-occurrence Matrix (GLCM). A detailed description of these visual feature extractions is provided in Sec. V-B.

To process this high-dimensional input while mitigating overfitting and preserving relevant structure, HV-Net employs a convolutional autoencoder (CNN-AE) built on 1D convolutional layers. The use of 1D-CNNs, instead of 2D-CNNs, is motivated by the nature of the input, which is a flattened feature vector without spatial dimensions. This choice significantly reduces the number of trainable parameters, improves generalization, and allows the model to capture local patterns efficiently. The encoder consists of sequential 1D-CNN layers with filter sizes of 256, 256, 128, 64, and 32, and kernel sizes of (1×4), (1×4), (1×3), (1×3), and (1×3), respectively. Each convolutional layer is followed by a max pooling operation with a pooling size of (1×2) to reduce temporal resolution and improve robustness. The decoder mirrors this structure in reverse order, applying 1D-CNN layers with filter sizes of 32, 64, 128, 256, and 256 to reconstruct the original input feature vector. This self-supervised reconstruction allows the network to learn stable and discriminative visual features by suppressing irrelevant variations while retaining meaningful texture structure.

The output of the decoder is passed through two fully connected layers with 512 and 128 units, respectively, each followed by a ReLU activation. This projection compresses the learned representation into a compact 128-dimensional visual feature vector, which is later combined with the tactile features from HT-Net during the multimodal fusion stage.

B. HAPTIC TACTILE NETWORK (HT-NET)

During surface exploration, stroking motions involve user-controlled interaction parameters such as scanning speed (v) and applied force (f), which determine how the surface is explored. These interactions produce vibrations that reflect surface properties including microstructure, roughness, and friction. The resulting dynamic responses are captured through acceleration signals (a), which contain high-frequency components induced by contact with the surface. Acceleration signals are often affected by sensor

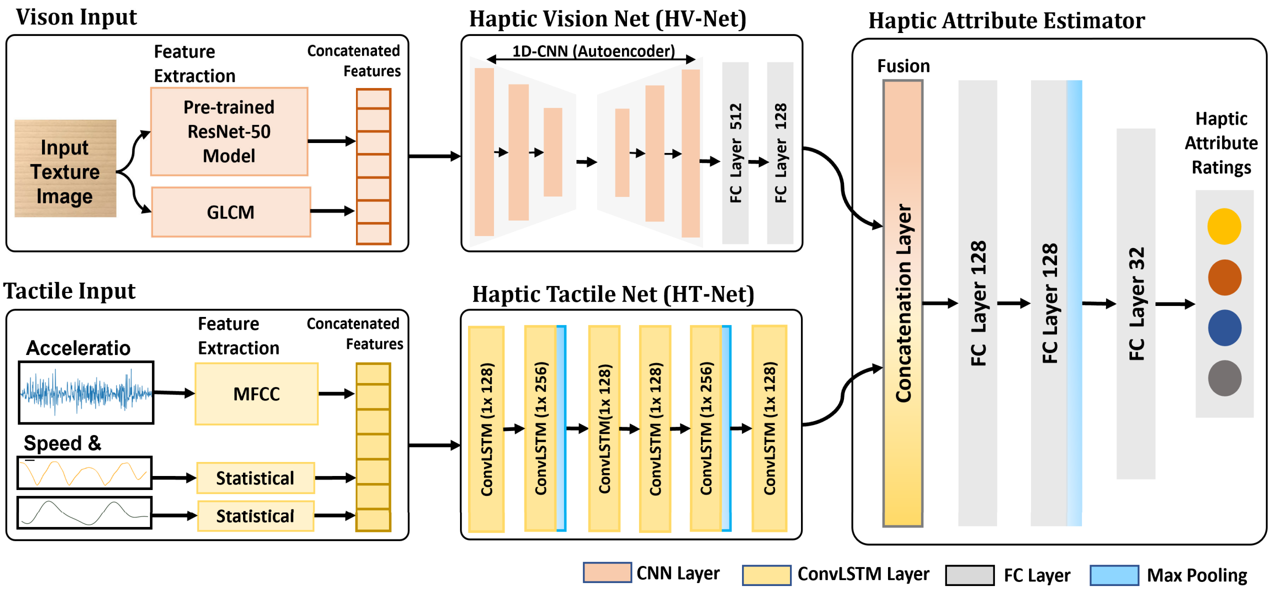

FIGURE 2. The proposed Visuo-Tactile Net. It consists of two streams: one for visual data employing an autoencoder and another for tactile data utilizing a 1D CNN.

noise and variability in hand motion, particularly under unconstrained conditions. To improve robustness, HT-Net operates on extracted features rather than raw signals. The continuous tactile recordings are first segmented into overlapping temporal windows, with each segment treated as a single input instance. For the acceleration signal, Mel-Frequency Cepstral Coefficients (MFCCs) are computed to capture spectral content in a compact, noise-resilient form. In contrast, scanning speed and applied force are low-frequency signals that exhibit limited variation within short intervals. Therefore, statistical descriptors are computed from the speed and force signals to summarize their temporal behavior across each segment. The complete feature extraction process is described in Sec. V-A.

For each temporal segment, a feature vector is defined as

$$X_t = (\mathrm{MFCC}_a, \mathrm{statistical}(v), \mathrm{statistical}(f)),$$

where MFCC_a denotes cepstral features extracted from a, and statistical(v) and statistical(f) represent statistical descriptors computed separately from v and f. Each segment-level vector X_t forms one input in the sequence provided to the ConvLSTM. The MFCC features extracted from acceleration are concatenated with the statistical descriptors of scanning speed and applied force to form a unified feature vector per temporal segment. This fused vector serves as one time step in the sequence input to the HT-Net.

HT-Net uses a ConvLSTM-based architecture to effectively capture both local spatial structure and long-range temporal dependencies present in sequential tactile signals. ConvLSTM combines convolutional operations with recurrent memory, making it particularly suitable for spatio-temporal modeling of tactile data [17]. The network comprises six stacked 1D-ConvLSTM layers with filter sizes of 128, 256, 128, 128, 256, and 128, respectively. To reduce temporal resolution and enhance representational efficiency, temporal max pooling operations with a window size of 1 × 2 are applied after the 2nd, 4th, and 5th ConvLSTM layers. This layer-wise architecture enables HT-Net to extract hierarchical tactile representations while progressively compressing the temporal dimension. The final hidden state output forms a compact 128-dimensional tactile representation, which is later fused with the visual stream for joint haptic attribute prediction. Each ConvLSTM layer follows the original formulation introduced in [17] and is implemented using TensorFlow Keras [42].

C. OUTPUT AND TRAINING METHOD

The final visual and tactile representations from HV-Net and HT-Net, each 128-dimensional, are concatenated to form a 256-dimensional multimodal feature vector. This vector passes through two fully connected (FC) layers with 128 units, followed by a max pooling layer with a window size of 1 × 2. The pooled output is processed by an FC layer with 32 units and a final FC layer with 4 output neurons, which predict the haptic attribute scores.

The configuration of the overall architecture, including the number of layers, filter sizes, and fully connected dimensions, was determined through extensive empirical experiments. The network is trained end-to-end using the TensorFlow-Keras framework with the Adam optimizer and RMSE loss. ReLU activation is used in all intermediate layers, while the final output layer employs linear activation to support continuous regression. Training is performed for up to 200 epochs, with early stopping based on validation performance and a patience of 10 epochs.

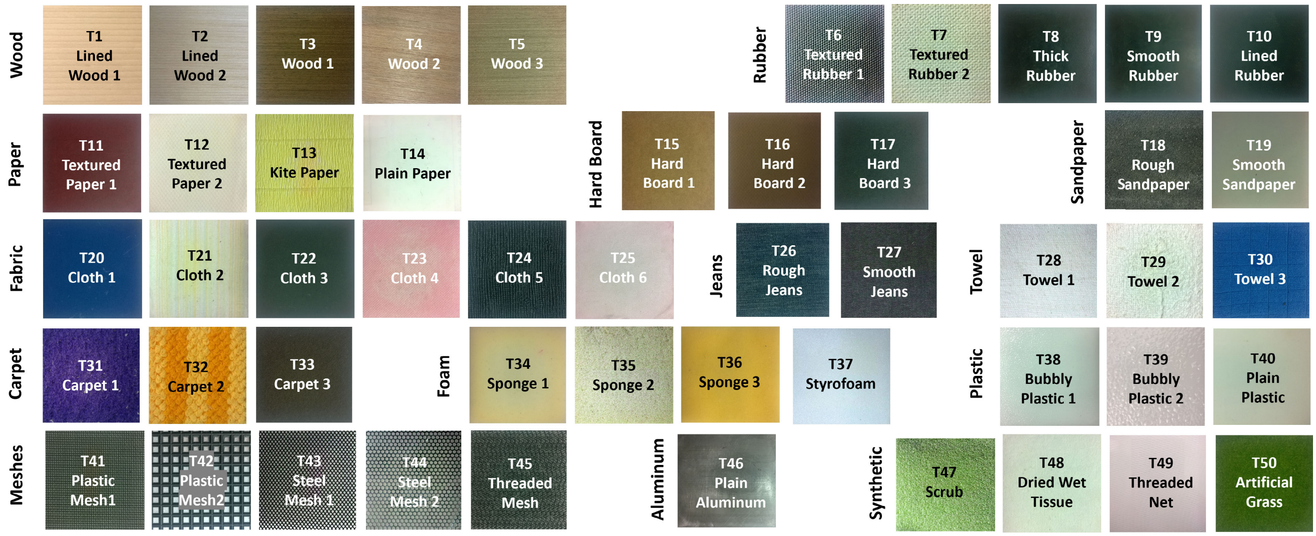

FIGURE 3. The real texture samples used in this study from diverse categories.

IV. HAPTIC PERCEPTUAL SPACE (HPS)

This section briefly describes the process of creating the proposed Haptic Perceptual Space (HPS) using human participants. The first step involves conducting a psychophysical experiment to identify texture attributes that characterize perceptual properties. In the second part of the experiment, participants rated the texture attributes identified in the prior phase. It is noted that the attribute rating dataset and experimental setup used for perceptual evaluation were adopted from our previous study [15], while the visual and tactile dataset described in Section V was newly collected in this work. In addition, the current study introduces a novel structuring and visualization of the perceptual ratings in a four-dimensional attribute space to reveal the spatial distribution of textures based on actual user ratings.

A. TEXTURE DATASET

This study utilizes 50 real texture samples from diverse categories to construct both the Haptic Perceptual Space (HPS) and the Physical Signal Space (PSS). The textures were carefully selected to represent a broad spectrum of materials and surface properties. To ensure comprehensive coverage, the dataset was categorized into 16 distinct classes. Each class contains textures exhibiting various characteristics, including differences in roughness, softness, slipperiness, and other tactile properties. The texture categories include wood, rubber, paper, hardboard, sandpaper, fabric, jeans, towels, carpet, foam, plastic, meshes, aluminum, and synthetic materials. A detailed overview of the 50 texture samples is presented in Figure 3.

Furthermore, all the texture samples were cut to 100 × 100 mm for standardization. They were then affixed to pre-prepared hard acrylic plates of the same dimensions. The acrylic plates, measuring 100 × 100×5 mm, ensured uniform surface elevation across all samples. Liquid surface glue was used to attach the textures securely. This mounting process was implemented to ensure uniform surface elevation across all samples and to prevent any unevenness or curling of the texture during the experiments [14], [15].

B. EXPERIMENT 1: ATTRIBUTE SELECTION

The initial phase of this psychophysical study focused on identifying key adjectives that describe human perception of surface textures upon interaction. For this study, we gathered 60 haptic texture-related attributes/adjectives that can represent the dataset. A total of three sources were used to gather these adjectives: literature [1], [3], domain knowledge, and user experiments. The full list of adjectives used during the experiment is shown in Table 1.

TABLE 1. Attributes presented to participants for the attribute selection experiment. The four selected attribute pairs, highlighted in bold dark blue, were subsequently used for the rating experiment.

Refined	Jarred	Bald	Mushy	**Flat**	Vague
Furry	Grating	Silky	Warm	Thick	**Smooth**
Hard	Bouncy	Pleasant	Glassy	Pointy	Blur
Sticky	Sharp	Dense	Angular	Hatched	Even
Jagged	Spongy	**Bumpy**	Cold	Slow	Dark
Grainy	Patterned	**Slippery**	Light	Slick	Granular
Distinct	Irritating	Wooden	Mild	Bright	**Rough**
Prickly	Metallic	Bubbly	Deep	Fast	Heavy
Solid	Fine	Blur	Shallow	Rigid	**Soft**
Glassy	Thin	Hatched	Sparse	Blunt	Fizzy

1) PARTICIPANTS

A total of 26 participants (19 male and 7 female) took part in this study, with ages ranging from 25 to 34 and an average age of 28. All participants were right-handed, used their dominant hand during the experiment, and reported no disabilities that might impact their performance or require special accommodations.

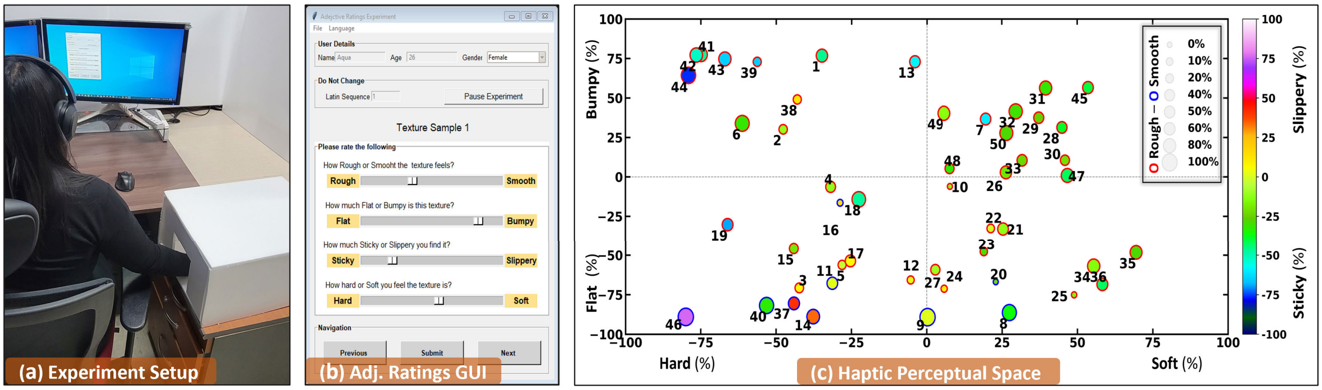

FIGURE 4. (a) The perceptual experiment setup. (b) The GUI for adjective ratings experiment. (c) Four-dimensional haptic perceptual space visualized as a 2D bubble plot. The plot shows ratings from four adjective pairs for each texture: hard-soft (x-axis), flat-bumpy (y-axis), rough-smooth (bubble size), and sticky-slippery (color gradient). Ratings range from -100 to 100, with -100 representing one extreme (e.g., hard) and 100 the opposite (e.g., soft).

2) EXPERIMENT SETUP

The experimental setup is illustrated in Figure 4. Participants were seated at a table, wearing headphones that emitted white noise to minimize environmental distractions. A cardboard box with two openings was placed on the table. One opening featured a small aperture through which participants could insert their hand to explore the textures, effectively blocking visual input during the task. The second opening allowed the experimenter to interchange the textures without revealing them to the participant. Participants received instructions in both written and verbal form, detailing the task of selecting adjectives to describe the perceived textures.

3) STIMULI AND PROCEDURE

The primary objective of this experiment is to identify adjectives that characterize the perception of texture. Each participant was presented with 50 texture samples (see Fig. 3), one at a time, and allowed to explore them freely through touch without time constraints, using any preferred exploratory movements. Participants evaluated each texture individually and selected adjectives from the provided list (see Table 1) that they considered relevant. Their decisions were recorded in binary form: "1" for relevant adjectives and "0" for irrelevant ones.

4) RESULTS

The analysis revealed key attributes that consistently described the texture surfaces. The scores assigned to each adjective across all textures and participants were summed and normalized to generate a relevance score. Adjectives with relevance scores of 50% or higher were retained for further analysis, yielding a selection of 11 adjectives. From this set, antonymous pairs were identified to represent opposing ends of perceptual dimensions. Adjectives without corresponding antonyms were excluded. The final set consisted of four antonymous pairs: rough–smooth, flat–bumpy, sticky–slippery, and hard–soft.These pairs were used in the next phase of the experiment.

C. EXPERIMENT 2: ADJECTIVE RATINGS

1) EXPERIMENT SETUP

In the second phase, participants rated each texture using antonymous attribute pairs identified in the first experiment. The ratings were recorded through a custom-designed user interface displayed on a PC (see Fig. 4). This interface featured four sliders, each representing one of the antonymous pairs. The physical length of each slider was 127 mm, following the standardized method [43]. This design ensured sufficient resolution for participants to express subtle perceptual differences, enhancing the precision of data collection in perceptual scaling experiments while maintaining ease of use [43], [44]. Participants explored each texture with their dominant hand, taking as much time as needed to reach a confident assessment. Slider values ranged from 0 to 100, with each slider representing a scale between two opposing attributes displayed at either end, while the numerical values remained hidden from participants.

2) RESULTS

The responses from all participants were aggregated to derive the final perceptual ratings for each texture. For enhanced analysis and visualization, these ratings were averaged and mapped onto a scale ranging from -100 to 100, with 0 representing the midpoint. On this scale, -100 and 100 correspond to the extremes of each attribute (e.g., rough to smooth), with polarity indicating the shift toward opposing haptic properties.

The final outcome of this study is the average rating for each attribute corresponding to each texture, which will be used to map physical signal space to perceptual space, as described in Section III. To visualize this four-dimensional dataset, we developed a haptic perceptual space (HPS) using a bubble chart with a color gradient. The HPS encodes four dimensions: hard-soft (x-axis), flat-bumpy (y-axis), rough-smooth (bubble size), and sticky-slippery (color gradient). The HPS plot is illustrated in Fig. 4. To the best of our knowledge, this HPS is the first visualization

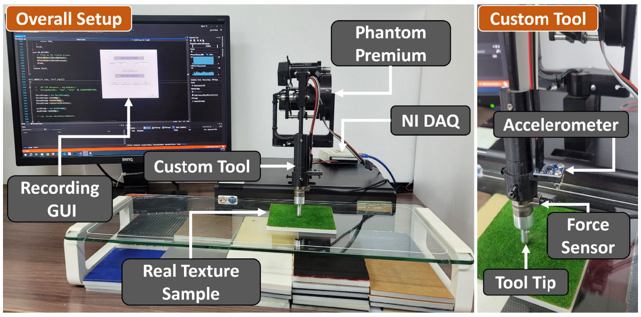

FIGURE 5. Data recording setup. The setup records vibrations produced when the user rubs the surface, along with the applied speed and force.

to consolidate multi-dimensional haptic attributes into a unified 2D framework to display absolute ratings. Unlike previous studies that required multiple graphs to represent each dimension separately [15], this approach integrates all sensory dimensions within a single plot, streamlining the interpretation of texture properties and enabling efficient analysis of large datasets.

V. PHYSICAL FEATURE SPACE (PFS)

This section defines the Physical Feature Space (PFS), a dataset composed of synchronized tactile signals and visual observations. The first part describes the tactile data, including the hardware configuration, signal acquisition, preprocessing, and feature computation. The second part outlines the visual data, covering image capture and the extraction of both deep and classical texture descriptors.

A. TACTILE DATASET

1) APPARATUS

The tactile data acquisition setup is illustrated in Figure 5. It consists of a rigid tool equipped with a detachable 2.0 mm hemispherical stainless steel tip. The tool body is custom-designed and fabricated using ABS plastic. A 3-axis accelerometer (ADXL335, Analog Devices) is mounted on the tool to record vibrations during surface exploration, while a force sensor (Nano17, ATI Industrial Automation) measures forces along three axes. The tool is mounted on a Phantom Premium haptic device, enabling precise tracking of position and orientation for accurate speed at 1 kHz and normal force estimation. The accelerometer connects to a PC via a data acquisition card (USB-6351, National Instruments), recording at 3 kHz. The force sensor uses a dedicated DAQ system to sample forces at 8 kHz. This hardware configuration is consistent with setups commonly used in haptic research for texture data collection and offers high-resolution measurements suitable for tactile signal analysis [45].

2) DATA COLLECTION AND PRE-PROCESSING

Interaction data was collected for all 50 textures detailed in Sec. IV-A. Each texture was recorded for 60 seconds using freehand motion to capture natural surface interactions.

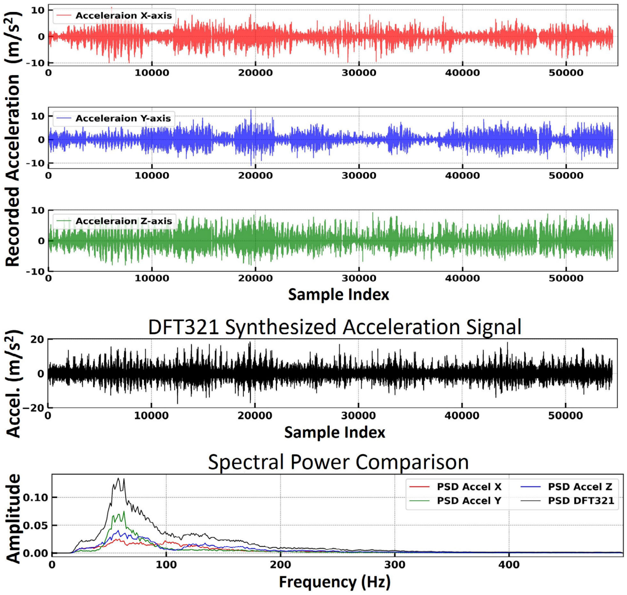

FIGURE 6. Acceleration signals for artificial grass recorded along three axes (first three plots) and combined into a single-axis using the DFT321 algorithm (fourth plot), retaining temporal characteristics and spectral power (fifth plot).

All recorded data was resampled at 1000 Hz for uniformity. The initial and final 2.5 seconds were cropped to reduce artifacts and eliminate stationary effects. Interaction signals, including scanning speed and normal force, were low-pass filtered at 25 Hz to suppress high-frequency noise, while acceleration signals were band-pass filtered between 20 Hz and 500 Hz to isolate relevant vibrations and remove gravitational components [6], [45]. The 3-axis acceleration signals were projected onto a single axis using the DFT321 algorithm, preserving temporal and spectral characteristics [46]. Scanning speed was derived by combining velocities along all three axes, and normal force was computed by projecting 3-axis force vectors onto the surface normal. Figure 6 shows the final processed data for artificial grass texture (T50).

3) MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCC)

Physical acceleration signals collected from textured surfaces contain valuable haptic information alongside redundant data. To extract meaningful haptic features from these signals, we apply Mel Frequency Cepstral Coefficients (MFCC), a technique widely used in audio and signal processing [47]. MFCC effectively captures essential vibrational patterns from texture data, making it well-suited for haptic analysis. It has been successfully applied to surface classification and texture modeling [28], [29].

To compute MFCCs, the raw acceleration signals were segmented using a sliding window approach with 50% overlap, which effectively increases the number of training samples available to the model while preserving temporal dynamics. Specifically, 0.5-second segments (500 samples at 1000 Hz) were extracted using a Hann window, with each

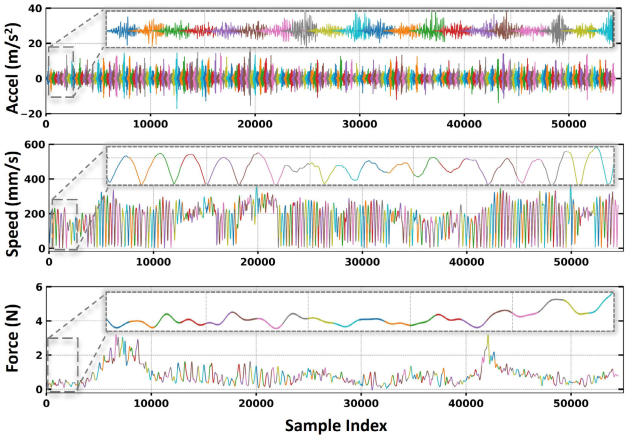

FIGURE 7. The processed acceleration signal along with interaction signals, including speed and force. Each graph also shows the segments created during pre-processing.

segment overlapping the previous one by 250 samples (see Figure 7). Each segment was further divided into 25 ms frames with 50% internal overlap, resulting in 40 frames per segment. For each frame, 13 MFCC coefficients were extracted, producing a matrix of size 40 × 13. Flattening this matrix yielded 520 MFCC features per segment.

On the other hand, speed and force signals are low-frequency and exhibit minimal variation over short intervals; their minimum, maximum, and average values were computed for each segment, contributing 6 additional features. The final feature vector, comprising 526 features per segment, was computed using Python's SciPy and librosa libraries. Each segment's features serve as input to the tactile network (HT-Net). An illustration of the segmented acceleration, speed, and force signals for Artificial Grass surface (T50) is shown in Figure 7.

B. IMAGE DATASET

The proposed multi-modal strategy incorporates texture images to extract visual features for predicting haptic ratings. Traditionally, haptic applications have used raw images with classical texture descriptors like Gray-Level Co-occurrence Matrix (GLCM), Gabor filters, and Local Binary Patterns (LBP) [12]. Recently, pre-trained deep learning models have become popular for visual feature extraction in texture analysis and tactile perception [35]. Despite their effectiveness, deep learning models can miss surface details when target textures differ from those in the training datasets. To mitigate this, we adopted a hybrid approach combining classical and deep learning-based feature extraction. GLCM, a robust texture descriptor, was used alongside features from a pre-trained deep learning model.

1) IMAGE CAPTURING SETUP

Learning haptic properties from images requires capturing fine surface details and granularity. To achieve this, high-resolution images are essential. We developed a setup using a dp2 Quattro SIGMA camera mounted on a tripod, maintaining a fixed distance of 30 cm between the camera lens and the surface. For each of the 50 textures, 10 images were captured under varying lighting and angular conditions to enhance feature diversity and generalization. To minimize boundary blur, all images were cropped and resized to 1568 X 1568 from the center.

2) DL-BASED FEATURES

ResNet [16], known for its deep architecture and residual connections, has demonstrated significant performance in image classification and feature extraction tasks. Its ability to capture fine-grained details makes it well-suited for applications requiring dense visual representations, including haptic texture analysis [35].

In this study, we employed ResNet-50 [16], pre-trained on ImageNet, to extract feature vectors, a method validated in prior haptic research [15], [35]. To maintain the resolution of texture images and avoid loss of detail, each image was divided into 49 overlapping 224 × 224 patches, matching ResNet's input size. For each patch, feature vectors of size 1 × 2048 were extracted from the average pooling layer. These vectors were averaged across all patches to generate the final feature representation for each image. Notably, since texture perception is less dependent on color, all images were converted to grayscale. To ensure compatibility with ResNet's three-channel input, grayscale images were replicated across three channels, allowing seamless use of the pre-trained model without modifying its architecture.

3) CLASSICAL TEXTURE DESCRIPTORS

For classical texture analysis, we employed the Gray-Level Co-occurrence Matrix (GLCM) [48], a widely used method for capturing spatial relationships between pixel intensities. GLCM has also been extensively applied in texture analysis and has shown significant success in haptic studies for characterizing surface properties [15]. In this study, the GLCM was computed on surface texture images quantized to 16 gray levels, resulting in a 16 × 16 matrix. This matrix was then flattened to generate a feature vector of size 1 × 256.

In the final stage, the extracted GLCM features were combined with deep learning (DL)-based features obtained from ResNet-50. The 2048-dimensional feature vectors were extracted from the average pooling layer of ResNet-50 and concatenated with the GLCM features, resulting in a comprehensive image-based feature vector of size 1 × 2304. This combined feature vector served as the input to the Haptic Vision Network (HV-Net) for further processing. It is noted that, to improve generalization and better align the visual input space with the temporal tactile segments, visual samples were dynamically augmented during training using TensorFlow's data pipeline. The augmentation process included random rotations, horizontal and vertical flips, and Gaussian noise, thereby enriching visual diversity across input conditions.

VI. EVALUATION EXPERIMENTS

The primary objective of this experiment is to evaluate the effectiveness of the proposed approach in estimating perceptual attributes of textured surfaces. The following sections outline the error metrics, the leave-one-out cross-validation (LOOCV) technique for unseen data, and the results obtained. The framework is compared with existing methods, followed by an analysis of different feature sets.

A. ERROR METRICS

To evaluate the performance of the proposed framework, we employed Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) as the primary error metrics. MAE quantifies the average magnitude of errors, while RMSE penalizes larger deviations, providing a comprehensive measure of prediction accuracy. These metrics were used to assess the model's effectiveness in estimating individual haptic attribute estimation accuracy by comparing predicted values with user-provided ratings. The attributes evaluated are the same as those discussed in Sec. IV, including Rough-smooth (R-S), Flat-bumpy (F-B), Sticky-slippery (S-S), and Hard-soft (H-S). These metrics are widely used in related studies [14], [15] and are defined as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \tilde{y}_i|, \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2}, \tag{2}$$

where y_i represents the actual rating provided by the user for the i^{th} texture sample, $\tilde{y}_i$ denotes the estimated attribute rating, and n is the total number of observations or texture samples. It is important to note that the actual and predicted values are scaled to the range of 0 to 100 before computing the error. For example, an MAE of 10 represents an average deviation of 10 on a 100-point scale, reflecting the difference between predicted and actual user ratings.

B. LEAVE-ONE-OUT CROSS VALIDATION (LOOCV)

Fitting high-dimensional data requires robust validation techniques to identify the most optimized models and ensure reliable performance across different methods. One of the most widely used validation approaches is cross-validation, which evaluates a model's ability to generalize by repeatedly training and testing it on different subsets of data [49]. Among the various types of cross-validation, k-fold cross-validation is the most common. In this approach, the dataset is divided into k equally sized subsets (typically $k = 5$ or $k = 10$). During each iteration, one subset is held out for validation, while the remaining $k - 1$ subsets are used for training. This process repeats k times, with each subset serving as the validation set once. The model's final performance is averaged across all iterations. However, choosing a small k (such as 5 or 10) can sometimes lead to high bias and underfitting, particularly when dealing with small datasets. To address this, leave-one-out cross-validation (LOOCV), a special case of k-fold cross-validation where $k = n$ (the total number of observations), is often employed. In LOOCV, the data is split into n subsets, where each iteration uses a single data point for validation while the remaining $n - 1$ points are used for training. This process repeats for every observation, ensuring that each data point is tested exactly once, which significantly reduces bias and leverages the entire dataset for model training [49], [50], [51].

In this work, we utilized LOOCV to evaluate the performance of the proposed framework. Our dataset consists of 50 textures ($n = 50$). For each iteration, 49 textures ($n - 1$) were used for training, while the remaining texture was reserved for validation. This resulted in 50 training cycles, providing a comprehensive assessment of the model's generalizability. Despite being computationally demanding, LOOCV is particularly valuable for small datasets, as it maximizes the use of available data and yields reliable performance estimates on unseen samples. This makes it an ideal choice for our evaluation process.

C. MODEL PERFORMANCE

Figure 8 presents the comparison between actual and predicted values from the proposed visuo-tactile Net for each texture. The results are plotted within a range of -100 to 100 for each surface. As shown, the model's predictions align closely with user ratings for most textures.

To further assess accuracy, we calculated MAE across all attributes. The lowest error of 4.48 was recorded for the F-B attribute, followed by H-S and R-S with 5.21 and 5.23, respectively. The highest error of 6.67 was observed for the S-S attribute, as shown in Table 2. All MAE values are scaled from 0 to 100, as described in Sec. VI-A, ensuring consistent interpretation across different attributes. Additionally, class-based errors are visualized in Figure 9. The results show that paper and jeans textures exhibited the highest errors, whereas other texture classes achieved stronger predictive performance. Further details are provided in Sec. VII.

D. COMPARISON WITH BASELINE MODELS

The performance of the proposed framework was evaluated against other similar strategies for estimating haptic texture attribute ratings using either visual and/or tactile data. These include TactResNet [52], Haptic CNN [15], Tactile CNN-LSTM [14], Tactile SVM [53], and a multimodal artificial neural network (ANN) baseline. All models were implemented using TensorFlow 2.7, and their core architectures were reproduced based on the original publications. For consistency, the final regression layer of each model was modified to produce four continuous outputs corresponding to the four haptic attributes. The ANN baseline provides a simplified multimodal fusion benchmark without explicit

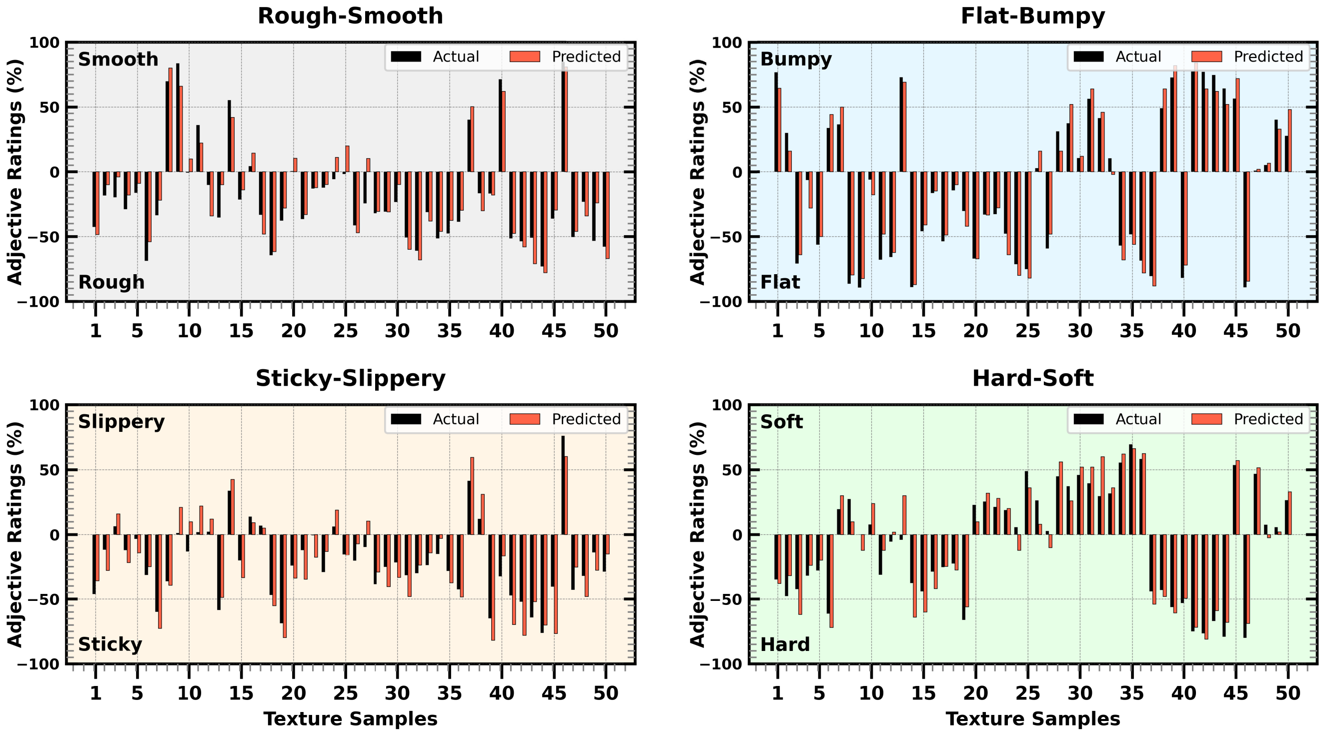

FIGURE 8. Comparison of actual and predicted attributes for 50 textures using the Leave-One-Out Cross-Validation (LOOCV) technique.

TABLE 2. Mean Absolute Error (MAE) values for the proposed system and five algorithms across four attribute pairs.

Methods	R-S	F-B	S-S	H-S
Artificial Neural Network	21.13	26.12	22.85	25.44
TactResNet [52]	15.78	15.83	17.21	16.54
Haptic CNN [15]	13.17	11.32	12.01	8.38
Tactile CNN-LSTM [14]	10.58	8.98	13.76	11.92
Tactile SVM [53]	9.40	14.89	15.35	10.54
Proposed Method	**5.23**	**4.48**	**6.67**	**5.21**

modeling of spatial or temporal structure. It uses the same extracted features as the proposed framework: flattened ResNet and GLCM features for vision, and MFCC with statistical descriptors for tactile input. These are passed through two parallel fully connected branches (layer sizes: 128, 256, 256, 128), followed by feature fusion and two regression layers of 64 units. A final dense layer outputs four predicted values. This setup allows examination of the benefits introduced by modality-specific and structured processing.

The results, shown in Table 2 for MAE and Table 3 for RMSE, demonstrate that the proposed method consistently outperforms baseline models across all attribute pairs. The proposed model achieved the lowest errors in both MAE and RMSE, reflecting its superior accuracy and generalizability. For MAE, the proposed method recorded values of 5.23 for R-S, 4.48 for F-B, 6.67 for S-S, and 5.21 for H-S. In contrast, the ANN exhibited significantly higher errors, with 21.13 for R-S and 25.44 for H-S. Similar trends were observed in RMSE, where the proposed model achieved the lowest errors at 6.81 (R-S), 5.67 (F-B), 7.52 (S-S), and 6.13 (H-S). The ANN, by comparison, yielded RMSE values of 24.41 (R-S) and 29.12 (H-S).

Among the baseline models, Tactile SVM [53] and CNN-LSTM [14] outperformed ANN but remained less accurate than the proposed method. For the F-B attribute, the proposed model achieved a lower MAE of 4.48 compared to 15.83 from TactResNet [52]. A similar trend appeared in RMSE, where [52] produced an error of 21.42, while the proposed method achieved a significantly lower RMSE of 5.67. Interestingly, vision-based models [15] and [52] consistently produced higher errors compared to tactile-based approaches, highlighting the advantage of tactile data for haptic attribute estimation and the strength of the proposed visuo-tactile multi-model based technique in outperforming vision-based or tactile-based approaches.

E. INDIVIDUAL FEATURE ERROR

In this section, we evaluate the performance of individual feature extraction techniques for both visual and tactile data, as well as the benefits of combining them. The goal is to identify which features contribute most to reducing errors in

TABLE 3. Root mean square error (RMSE) values for the proposed system and five algorithms across four attribute pairs.

Methods	R-S	F-B	S-S	H-S
Artificial Neural Network	24.41	31.62	25.73	32.19
TactResNet [52]	17.33	21.42	20.15	19.36
Haptic CNN [15]	18.21	12.15	14.19	12.65
Tactile CNN-LSTM [14]	13.45	10.65	15.20	13.78
Tactile SVM [53]	11.26	16.37	20.81	11.93
Proposed Method	**6.81**	**5.67**	**7.52**	**6.13**

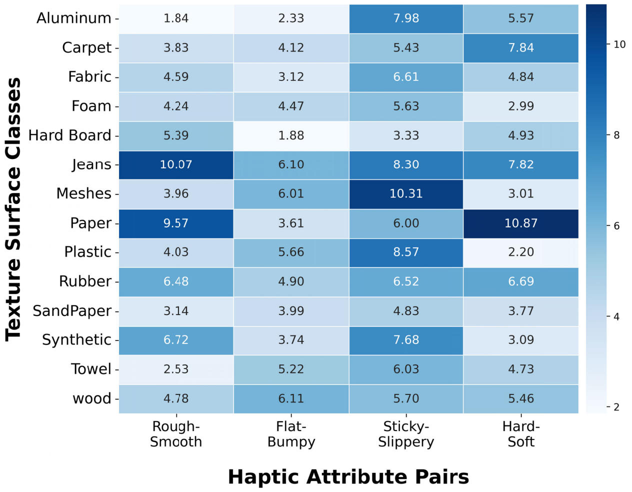

FIGURE 9. Heatmap of MAE for four haptic attribute pairs across various texture classes. Darker shades indicate higher errors, while lighter shades show lower errors, reflecting prediction performance across different texture classes using the visuo-tactile Net.

haptic attribute estimation. We assess features from ResNet-50 and GLCM for vision [15], [30], and 1D Discrete Wavelet Transform (1D-DWT), Discrete Fourier Transform (DFT), and MFCC for tactile data. These features were selected based on their effectiveness in haptic contexts [12], [31], [54].

Table 4 presents the performance of individual and combined features across both visual and tactile modalities. For vision-based inputs, concatenating ResNet and GLCM features led to improved accuracy across all attributes. The combined visual features achieved an RMSE of 10.11 for R-S, outperforming ResNet (18.29) and GLCM (19.11) individually. Similar improvements were observed for F-B and S-S. On the tactile side, MFCC consistently outperformed 1D-DWT and DFT, achieving an RMSE of 9.89 for R-S compared to 31.3 and 34.61, respectively. Notably, due to the poor performance of DWT and DFT, their combination with MFCC was not pursued, as initial trials led to unstable results and degraded performance. Combining visual and tactile features further reduced errors, resulting in the lowest RMSE across most attributes. The proposed model, integrating ResNet, GLCM, and MFCC, achieved RMSE values of 6.81 for R-S and 5.67 for F-B. Tactile data alone (MFCC, 11.35) also outperformed vision-only features for F-B, emphasizing the importance of tactile input for certain perceptual dimensions. Overall, the findings demonstrate the effectiveness of multi-feature, multimodal fusion in improving haptic attribute prediction.

TABLE 4. RMSE of individual features compared to concatenated features.

Feature Type	Feature	R-S	F-B	S-S	H-S
Vision	ResNet	18.29	16.52	15.36	13.50
	GLCM	19.11	12.53	10.14	14.96
	Concatenated	13.26	10.11	12.52	8.6
Tactile	1D-DWT	31.3	46.8	42.5	39.3
	DFT	34.61	29.85	26.51	28.41
	MFCC	9.89	11.35	10.71	7.98
Proposed Method	**ResNet+GLCM MFCC**	**6.81**	**5.67**	**7.52**	**6.13**

VII. DISCUSSION

Building on the findings presented in Figure 8 and Table 2, this section examines attribute-wise prediction trends, modality-specific behavior, and class-level error patterns to better understand the strengths and limitations of the proposed framework. Among the four attribute pairs, S-S exhibited the highest error, while F-B achieved the lowest, as reflected by the average MAE and RMSE. R-S and H-S showed moderate errors, performing better than S-S but not as accurately as F-B.

Notably, considering the effect of visual and tactile features, we found that each modality has its strengths, and their combination yields superior results. As shown in Table 4, the vision-based approach performed better in capturing the flat-bumpy (F-B) attribute compared to the tactile-based approach. This may be due to the visual features' ability to clearly detect surface patterns, while tactile signals, particularly acceleration data, may introduce noise during deep strokes, resulting in undesired bounciness.

Figure 9 highlights performance variations across texture classes, with paper and jeans categories exhibiting the highest errors across most attribute pairs. For paper textures, the highest MAE was recorded for H-S at 10.87 and R-S at 9.57, likely due to the diverse range of samples, including both plain and heavily textured surfaces. Since the model maps the physical signal space to a perceptual space derived from human ratings, it is plausible that participants may have overlooked finer details. Perceptual biases driven by preconceived judgments, as noted in [24], could have influenced ratings, where participants assess haptic qualities based on prior experiences rather than the actual textures presented during the experiment. A similar pattern emerged in the jeans category, particularly for T27 (smooth jeans), where surface texture variations likely contributed to increased errors, reflecting challenges akin to those encountered with paper textures.

Additionally, the meshes class showed elevated errors in the S-S attribute (MAE 10.31), which may be attributed to noise artifacts accumulating during tactile data recording. The rigid and structured nature of hard plastic and metal meshes could have introduced inconsistencies, resulting in higher prediction errors. Despite these discrepancies, the errors remain within acceptable bounds, aligning with the Just Noticeable Difference (JND) threshold for perceptual similarity, often estimated at around 10 out of 100 [38]. Most class-wise and overall average MAE values fall below this threshold, reinforcing the effectiveness of the proposed framework.

The generalizability of the model is further demonstrated by its performance on unique textures such as aluminum (T46), which exhibits distinct surface properties. Despite its uniqueness, aluminum performed well, with the highest error recorded in the sticky-slippery (S-S) attribute at 7.98. This elevated error may be attributed to rubbing marks left by the interaction tool, a known phenomenon in tactile studies. Since aluminum is the sole sample in its category, further investigation is necessary to better understand this behavior. We believe that incorporating additional textures with similar properties will enhance overall performance. However, it can be argued that the study has yet to encounter a sufficiently diverse range of textures. Expanding the dataset with additional textures is likely to improve texture attribute prediction quality. Although LOOCV can introduce biases in certain cases, it remains an effective method for comprehensive evaluation. The results clearly indicate that the proposed autoencoder-based framework, combined with CNN and feature-based inputs, captures nuanced surface properties and represents an improvement over existing single-modality approaches.

VIII. CONCLUSION

This study introduces a deep learning visuo-tactile framework for predicting haptic texture attributes. It maps a physical signal space, constructed from visual and tactile features, to a perceptual space defined by user ratings. The four-dimensional perceptual space includes the bipolar pairs: rough–smooth, flat–bumpy, hard–soft, and sticky–slippery. The architecture combines a CNN-based autoencoder for visual processing with a ConvLSTM network for modeling tactile signal dynamics. Visual inputs are encoded using features from ResNet and GLCM, while tactile signals are represented using MFCCs derived from high-frequency acceleration data. The framework demonstrates improved prediction accuracy over existing methods by integrating visual and tactile data in a unified manner. These results confirm the framework's reliability and scalability in estimating haptic attributes, with potential utility in material recognition when user ratings are unavailable or difficult to collect, assisting researchers in assessing perceptual responses, and in robotic perception systems where accurate surface interpretation is essential.

REFERENCES

[1] Y. Yoo, J. Lee, J. Seo, E. Lee, J. Lee, Y. Bae, D. Jung, and S. Choi, "Large-scale survey on adjectival representation of vibrotactile stimuli," in *Proc. HAPTICS*. New York, NY, USA: IEEE, 2016, pp. 393–395.

[2] B. A. Richardson, Y. Vardar, C. Wallraven, and K. J. Kuchenbecker, "Learning to feel textures: Predicting perceptual similarities from unconstrained finger-surface interactions," *IEEE Trans. Haptics*, vol. 15, no. 4, pp. 705–717, Oct. 2022.

[3] Y. Yoo, I. Hwang, and S. Choi, "Consonance of vibrotactile chords," *IEEE Trans. Haptics*, vol. 7, no. 1, pp. 3–13, Jan. 2014.

[4] K. Takahashi and J. Tan, "Deep visuo-tactile learning: Estimation of tactile properties from images," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 8951–8957.

[5] A. Abdulali and S. Jeon, "Data-driven rendering of anisotropic haptic textures," in *Haptic Interaction: Science, Engineering and Design 2*. Singapore: Springer, 2018, pp. 401–407.

[6] H. Culbertson, J. Unwin, and K. J. Kuchenbecker, "Modeling and rendering realistic textures from unconstrained tool-surface interactions," *IEEE Trans. Haptics*, vol. 7, no. 3, pp. 381–393, Jul. 2014.

[7] M. I. Awan, T. Ogay, W. Hassan, D. Ko, S. Kang, and S. Jeon, "Model-mediated teleoperation for remote haptic texture sharing: Initial study of online texture modeling and rendering," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 12457–12463.

[8] E. D. Gomez, H. M. Husin, K. R. Dumon, N. N. Williams, and K. J. Kuchenbecker, "Simulation training with haptic feedback of instrument vibrations reduces resident workload during live robot-assisted sleeve gastrectomy," *Surgical Endoscopy*, vol. 39, no. 3, pp. 1523–1535, Mar. 2025.

[9] S. Okamoto, H. Nagano, and Y. Yamada, "Psychophysical dimensions of tactile perception of textures," *IEEE Trans. Haptics*, vol. 6, no. 1, pp. 81–93, 1st Quart., 2013.

[10] K. Drewing, C. Weyel, H. Celebi, and D. Kaya, "Systematic relations between affective and sensory material dimensions in touch," *IEEE Trans. Haptics*, vol. 11, no. 4, pp. 611–622, Oct. 2018.

[11] P. Zhang, L. Bai, D. Shan, X. Wang, S. Li, W. Zou, and Z. Chen, "Visual–tactile fusion object classification method based on adaptive feature weighting," *Int. J. Adv. Robotic Syst.*, vol. 20, no. 4, Jul. 2023, Art. no. 17298806231191947.

[12] M. Strese, C. Schuwerk, A. Iepure, and E. Steinbach, "Multimodal feature-based surface material classification," *IEEE Trans. Haptics*, vol. 10, no. 2, pp. 226–239, Apr. 2017.

[13] D. Chen, D. Zhu, J. Liu, G. Chen, Y. Fang, and Y. Zhang, "Research on texture haptic reconstruction method based on informer model," in *Proc. 3rd Int. Conf. Robot. Control Eng.*, May 2023, pp. 161–165.

[14] M. I. Awan, W. Hassan, and S. Jeon, "Predicting perceptual haptic attributes of textured surface from tactile data based on deep CNN-LSTM network," in *Proc. 29th ACM Symp. Virtual Reality Softw. Technol.*, Oct. 2023, pp. 1–9.

[15] W. Hassan, J. B. Joolee, and S. Jeon, "Establishing haptic texture attribute space and predicting haptic attributes from image features using 1D-CNN," *Sci. Rep.*, vol. 13, no. 1, p. 11684, Jul. 2023.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Jun. 2016, pp. 770–778.

[17] X. Shi, Z. Chen, H. Wang, D. Yeung, W. K. Wong, and W. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2015, pp. 802–810.

[18] T. Yoshioka, S. J. Bensmaïa, J. C. Craig, and S. S. Hsiao, "Texture perception through direct and indirect touch: An analysis of perceptual space for tactile textures in two modes of exploration," *Somatosensory Motor Res.*, vol. 24, nos. 1–2, pp. 53–70, Jan. 2007.

[19] K. Yoshida, "The dimensions of tactile perception of surfaces," *J. Texture Stud.*, vol. 12, pp. 123–135, May 1968.

[20] M. Hollins, F. Lorenz, A. Seeger, and R. Taylor, "Factors contributing to the integration of textural qualities: Evidence from virtual surfaces," *Somatosensory Motor Res.*, vol. 22, no. 3, pp. 193–206, Jan. 2005.

[21] G. A. Gescheider, S. J. Bolanowski, T. C. Greenfield, and K. E. Brunette, "Perception of the tactile texture of raised-dot patterns: A multidimensional analysis," *Somatosensory Motor Res.*, vol. 22, no. 3, pp. 127–140, Jan. 2005.

[22] R. H. LaMotte, "Softness discrimination with a tool," *J. Neurophysiology*, vol. 83, no. 4, pp. 1777–1786, Apr. 2000.

[23] H. Culbertson and K. J. Kuchenbecker, "Ungrounded haptic augmented reality system for displaying roughness and friction," *IEEE/ASME Trans. Mechatronics*, vol. 22, no. 4, pp. 1839–1849, Aug. 2017.

[24] W. Hassan and S. Jeon, "Evaluating differences between bare-handed and tool-based interaction in perceptual space," in *Proc. IEEE Haptics Symp. (HAPTICS)*, Apr. 2016, pp. 185–191.

[25] E. Baumgartner, C. B. Wiebel, and K. R. Gegenfurtner, "Visual and haptic representations of material properties," *Multisensory Res.*, vol. 26, no. 5, pp. 429–455, 2013.

[26] V. Chu, I. McMahon, L. Riano, C. G. McDonald, Q. He, J. Martinez Perez-Tejada, M. Arrigo, T. Darrell, and K. J. Kuchenbecker, "Robotic learning of haptic adjectives through physical interaction," *Robot. Auto. Syst.*, vol. 63, pp. 279–292, Jan. 2015.

[27] J. Wu, N. Li, W. Liu, G. Song, and J. Zhang, "Experimental study on the perception characteristics of haptic texture by multidimensional scaling," *IEEE Trans. Haptics*, vol. 8, no. 4, pp. 410–420, Oct. 2015.

[28] W. Hassan, A. Abdulali, and S. Jeon, "Authoring new haptic textures based on interpolation of real textures in affective space," *IEEE Trans. Ind. Electron.*, vol. 67, no. 1, pp. 667–676, Jan. 2020.

[29] M. Strese, C. Schuwerk, and E. Steinbach, "Surface classification using acceleration signals recorded during human freehand movement," in *Proc. IEEE World Haptics Conf. (WHC)*, Jun. 2015, pp. 214–219.

[30] S. Lu, M. Zheng, M. C. Fontaine, S. Nikolaidis, and H. Culbertson, "Preference-driven texture modeling through interactive generation and search," *IEEE Trans. Haptics*, vol. 15, no. 3, pp. 508–520, Jul. 2022.

[31] N. Heravi, H. Culbertson, A. M. Okamura, and J. Bohg, "Development and evaluation of a learning-based model for real-time haptic texture rendering," *IEEE Trans. Haptics*, vol. 17, no. 4, pp. 705–716, Dec. 2024.

[32] F. Yang, C. Feng, Z. Chen, H. Park, D. Wang, Y. Dou, Z. Zeng, X. Chen, R. Gangopadhyay, A. Owens, and A. Wong, "Binding touch to everything: Learning unified multimodal tactile representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, vol. 33, Jun. 2024, pp. 26330–26343.

[33] S. Luo, W. Yuan, E. Adelson, A. G. Cohn, and R. Fuentes, "ViTac: Feature sharing between vision and tactile sensing for cloth texture recognition," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 2722–2727.

[34] W. Yuan, S. Dong, and E. Adelson, "GelSight: high-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, Nov. 2017.

[35] Z. Lin, H. Zheng, Y. Lu, J. Zhang, G. Chai, and G. Zuo, "Object surface roughness/texture recognition using machine vision enables for human-machine haptic interaction," *Frontiers Comput. Sci.*, vol. 6, May 2024, Art. no. 1401560.

[36] H. Li and H. Zhang, "Classification method of visual-tactile fusion dataset based on CNN-TCN," in *Proc. 8th Int. Conf. Control, Robot. Cybern. (CRC)*, Dec. 2023, pp. 295–300.

[37] W. Byeon, M. Liwicki, and T. M. Breuel, "Texture classification using 2D LSTM networks," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 1144–1149.

[38] M. I. Awan, A. Raza, W. Hassan, K.-U. Kyung, and S. Jeon, "Quantifying haptic affection of car door through data-driven analysis of force profile," 2024, *arXiv:2411.11382*.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.

[40] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, and L. Sun, "Transformers in time series: A survey," 2022, *arXiv:2202.07125*.

[41] Y. Zhang, Z. Kan, Y. Alexander Tse, Y. Yang, and M. Yu Wang, "FingerVision tactile sensor design and slip detection using convolutional LSTM network," 2018, *arXiv:1810.02653*.

[42] M. Abadi. (2015). *Tensorflow and Keras: Open-Source Deep Learning Frameworks*. Accessed: Mar. 13, 2024. [Online]. Available: https://www.tensorflow.org/

[43] S. Schiffman, *Introduction To Multidimensional Scaling: Theory, Methods, and Applications*. New York, NY, USA: Academic, 1981.

[44] I. Hwang and S. Choi, "Perceptual space and adjective rating of sinusoidal vibrations perceived via mobile device," in *Proc. IEEE Haptics Symp.*, Mar. 2010, pp. 1–8.

[45] A. Abdulali and S. Jeon, "Data-driven modeling of anisotropic haptic textures: Data segmentation and interpolation," in *Proc. Int. Conf. Human Haptic Sens. Touch Enabled Comput. Appl.* Cham, Switzerland: Springer, Jan. 2016, pp. 228–239.

[46] N. Landin, J. M. Romano, W. McMahan, and K. J. Kuchenbecker, "Dimensional reduction of high-frequency accelerations for haptic rendering," in *Proc. Haptics, Generating Perceiving Tangible Sensations, Int. Conf., EuroHaptics*, Amsterdam. Berlin, Germany: Springer, Jan. 2010, pp. 79–86.

[47] H.-G. Kim, N. Moreau, and T. Sikora, *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*. Hoboken, NJ, USA: Wiley, 2006.

[48] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[49] P. Patil, Y. Wei, A. Rinaldo, and R. J. Tibshirani, "Uniform consistency of cross-validation estimators for high-dimensional ridge regression," in *Proc. Int. Conf. Artif. Intell. Statist.*, Mar. 2021, pp. 3178–3186.

[50] V. Lumumba, D. Kiprotich, M. Mpaine, N. Makena, and M. Kavita, "Comparative analysis of cross-validation techniques: LOOCV, K-folds cross-validation, and repeated K-folds cross-validation in machine learning models," *Amer. J. Theor. Appl. Statist.*, vol. 13, no. 5, pp. 127–137, Oct. 2024.

[51] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *J. Roy. Stat. Soc. Ser. B: Stat. Methodology*, vol. 36, no. 2, pp. 111–133, Jan. 1974.

[52] J. M. Gandarias, A. J. García-Cerezo, and J. M. Gómez-de-Gabriel, "CNN-based methods for object recognition with high-resolution tactile sensors," *IEEE Sensors J.*, vol. 19, no. 16, pp. 6872–6882, Aug. 2019.

[53] Z. Shao, J. Bao, J. Li, and H. Tang, "Haptic recognition of texture surfaces using semi-supervised feature learning based on sparse representation," *Cognit. Comput.*, vol. 15, no. 5, pp. 1656–1671, Sep. 2023.

[54] A. Slepyan, M. Zakariaie, T. Tran, and N. Thakor, "Wavelet transforms significantly sparsify and compress tactile interactions," *Sensors*, vol. 24, no. 13, p. 4243, Jun. 2024.

MUDASSIR IBRAHIM AWAN (Graduate Student Member, IEEE) received the B.E. degree in electronics engineering from Karachi Institute of Economics and Technology (KIET), Karachi, Pakistan, in 2016. He is currently pursuing the integrated M.S. and Ph.D. degrees with the Department of Computer Science and Engineering, Haptics and Virtual Reality Laboratory, Kyung Hee University, South Korea. His research interests include data-driven modeling and rendering of haptic signals, psychophysical evaluation of haptic interfaces, and perception of tactile and kinesthetic stimuli.

SEOKHEE JEON received the B.S. and Ph.D. degrees in computer science and engineering from Pohang University of Science and Technology (POSTECH), in 2003 and 2010, respectively. He then worked as a Postdoctoral Research Associate with the Computer Vision Laboratory, ETH Zurich. In 2012, he joined the Department of Computer Engineering, Kyung Hee University, as an Assistant Professor and became a Full Professor, in 2024. He is also the Co-Founder and a Faculty Member with the Department of Metaverse. His research interests include data-driven haptic modeling and rendering, hyper-realistic multimodal feedback in virtual, augmented, and remote environments, and the development of modular wearable haptic interfaces with enhanced applicability.